Convex relaxation for IMSE optimal design in random field models

Bertrand Gauthier*† Luc Pronzato‡§

December 18, 2015

Abstract

The definition of an Integrated Mean-Squared Error (IMSE) criterion for the learning of a random field model yields a particular Karhunen-Loève expansion of the underlying field. The model can thus also be interpreted as a Bayesian (or regularised) linear model based on eigenfunctions of this Karhunen-Loève expansion, and can be approximated by a linear model involving orthogonal observation errors. Using the continuous relaxation of approximate design theory, the search of an IMSE optimal design can then be turned into a Bayesian A-optimal design problem, which can be efficiently solved by convex optimisation. We propose a greedy extraction procedure, of the exchange type, that permits to select observation locations among support points of an optimal design measure. In the presence of a parametric trend, we show how specific treatments can be applied to avoid confusion between the trend and eigenfunctions. The performance of the approach is investigated on a series of examples indicating that designs with very high IMSE-efficiency are easily obtained.

Keywords: random field model, Bayesian linear model, optimal design of experiments, integral operator, kernel reduction.

1 Introduction

This work addresses the problem of computing designs of experiments for second-order Random Field (RF) interpolation models with known covariance that are optimal in terms of Integrated Integrated Mean-Squared Error (IMSE), see, e.g., Sacks et al. (1989), Rasmussen and Williams (2006). The direct computation of IMSE-optimal designs for kernel-based models is often considered as a numerically challenging problem, see, e.g., Fang et al. (2010, Chap. 2), Santner et al. (2003, Chap. 6), in particular due to the presence of local minima.

The definition of an IMSE criterion for the leaning of a RF yields a particular Karhunen-Loève (KL) expansion of the considered field. Following Fedorov (1996); Spöck and Pilz (2010), for a given truncation level, we can interpret the initial RF model as a Bayesian (or regularised) Linear Model (BLM) based on a subset of eigenfunctions of the KL expansion. The IMSE criterion for this *exact* BLM corresponds to the truncated-IMSE criterion considered by Gauthier and Pronzato (2014).

*bgauthie@esat.kuleuven.be (corresponding author)
‡Luc.Pronzato@cnrs.fr
†KU Leuven, ESAT-STADIUS Center for Dynamical Systems, Signal Processing and Data Analytics
§CNRS, Laboratoire I3S – UMR 7271 Université de Nice-Sophia Antipolis/CNRS

However, the exact BLM involves correlated errors, and in order to apply the classical machinery of approximate design theory, we introduce an approximate BLM involving uncorrelated errors. Using continuous relaxation with design measures, the construction of an IMSE-optimal design is turned into a (convex) Bayesian A-optimal design problem, see Pilz (1983); Chaloner (1984) and Pukelsheim (1993, Chap. 11). Many convex optimisation algorithms are available to solve this problem, which ensure fast and guaranteed convergence to the optimum, see, e.g., Pronzato and Pázman (2013, Chap. 9).

We present a careful analysis of the approach and propose a numerical implementation based on a quadrature approximation of the IMSE. In particular, we investigate how to efficiently extract a design of given size n (the number of observations) from an optimal design measure, and how to construct an optimal measure adapted to the given n (notice that repeated observations at the same location are forbidden in an interpolation context).

For the sake of simplicity, we assume first (Sections 2 to 4) that the mean structure of the RF is known (and equal to zero without any loss of generality). The case of RF models involving an unknown linear parametric trend is considered in Section 6. Section 2 introduces the main notions and notation used in this work. In Section 3 we define the exact and approximate BLMs induced by the initial RF model and the IMSE criterion. In particular, we show the equivalence between the IMSE criterion of the exact BLM and the truncated-IMSE. We then consider two different choices for the variance structure of the observation noise in the approximate BLM: a homoscedastic model and a heteroscedastic model, both having the same integrated variance as the initial RF model. The Bayesian A-optimal design problem is introduced in Section 4. The numerical implementation of the approach is described in Section 5, assuming that a pointwise quadrature is used to approximate the integral of the Mean-Squared Error (MSE), and that the design space is restricted to quadrature points. The extraction of a design of given size n is also discussed. Section 6 describes how the convex relaxation approach can be applied to RF models with a linear parametric trend. We first recall the direct approach of Spöck and Pilz (2010). Next, depending on the presence of an informative prior on the first two moments of the trend-parameters, we propose two alternative ways for the construction of the exact BLM. When a prior is available, we consider the initial RF model with unknown trend as a RF model with known trend and augmented covariance kernel. In absence of prior, we consider a reduction of the initial kernel that brings orthogonality between the trend and the complementary centered RF, without any modification of the predictive properties of the model (see Theorem 6.1). Finally, some numerical experiments are carried out in Section 7, and Section 8 concludes.

2 General framework and notations

2.1 Random fields and related Hilbert structures

We consider a real RF $(Z_x)_{x\in\mathscr{X}}$ indexed by $\mathscr{X}$, where $\mathscr{X}$ can be any general set but corresponds to a compact subset of $\mathbb{R}^d$, $d \geqslant 1$, in most applications. In what follows Z will refer to the RF $(Z_x)_{x\in\mathscr{X}}$. We assume that Z is centered, second-order, and defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We denote by $L^2(\Omega, \mathbb{P})$ the Hilbert space of second-order real random variables (r.v.) on $(\Omega, \mathcal{F}, \mathbb{P})$, where we identify r.v. that are equal $\mathbb{P}$-almost surely. The inner product between two r.v. U and V of $L^2(\Omega, \mathbb{P})$ is denoted by $\mathsf{E}(UV)$.

Let $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ be the covariance kernel of Z, i.e., for all x and $y \in \mathscr{X}$, $\mathsf{E}(Z_x Z_y) = K(x, y)$.

Also, let $\mathbb{H}$ be the Hilbert space (sometimes called Gaussian Hilbert space) associated with Z, i.e., the closed linear subspace of $L^2(\Omega, \mathbb{P})$ spanned by the r.v. Z_x, $x \in \mathscr{X}$, endowed with the Hilbert structure induced by $L^2(\Omega, \mathbb{P})$. We assume that $\mathbb{H}$ is separable.

We denote by $\mathcal{H}$ the RKHS of real valued functions on $\mathscr{X}$ defined by the kernel $K(\cdot, \cdot)$. The two Hilbert spaces $\mathcal{H}$ and $\mathbb{H}$ are isometric thanks to the relation, for all x and $y \in \mathcal{X}$, $(K_x | K_y)_{\mathcal{H}} = K(x, y) = \mathsf{E}(Z_x Z_y)$, where $(\cdot|\cdot)_{\mathcal{H}}$ is the inner product of $\mathcal{H}$. We denote by $\mathcal{I} : \mathcal{H} \to \mathbb{H}$ the isometry given by $\mathcal{I}(K_x) = Z_x$, where K_x stands for the function $t \mapsto K(x, t)$, $t \in \mathscr{X}$.

2.2 Hilbert space embedding and integral operator

We suppose that $\mathscr{X}$ is a measurable space; we denote by $\mathcal{A}$ the associated σ-algebra and consider a σ-finite measure μ on $\mathscr{X}$ (this is the measure used to define the IMSE criterion, see Section 2.4). We denote by $L^2(\mathscr{X}, \mu)$ the Hilbert space of real-valued functions on $\mathscr{X}$ that are square integrable with respect to μ. Notice that elements of $L^2(\mathscr{X}, \mu)$ are in fact equivalent classes of functions that coincide μ-almost everywhere; however, we shall assimilate elements of $L^2(\mathscr{X}, \mu)$ with functions on $\mathscr{X}$ when it will not be source of confusion.

We assume that the kernel $K(\cdot, \cdot)$ is measurable on $\mathscr{X} \times \mathscr{X}$ endowed with the product σ-algebra, and that the diagonal of $K(\cdot, \cdot)$ is a measurable function on $\mathscr{X}$. We also assume that the RKHS $\mathcal{H}$ is continuously included into $L^2(\mathscr{X}, \mu)$, that is, for any $h \in \mathcal{H}$, we have $h \in L^2(\mathscr{X}, \mu)$ and

$$\|h\|_{L^2}^2 \leqslant \tau \|h\|_{\mathcal{H}}^2, \text{ with } \tau = \int_{\mathscr{X}} K(x, x) \mathrm{d}\mu(x) < +\infty;$$

one may refer for instance to Gauthier and Pronzato (2014) for more precisions. We denote by $\mathcal{H}_0$ the closed linear subspace of $\mathcal{H}$ defined by $\mathcal{H}_0 = \{h_0 \in \mathcal{H} | \, \|h_0\|_{L^2}^2 = 0\}$ and by $\mathcal{H}_\mu$ the orthogonal of $\mathcal{H}_0$ in $\mathcal{H}$ (i.e., $\mathcal{H}_\mu = \mathcal{H}_0^{\perp_{\mathcal{H}}}$).

We introduce the following linear operator T_μ on $L^2(\mathscr{X}, \mu)$,

$$\forall f \in L^2(\mathscr{X}, \mu), \quad \forall x \in \mathscr{X}, \quad T_\mu[f](x) = \big(K_x | f\big)_{L^2} = \int_{\mathscr{X}} f(t) K(x, t) \mathrm{d}\mu(t).$$

The operator T_μ is compact, positive and self-adjoint, and $T_\mu[f] \in \mathcal{H}_\mu$ for all $f \in L^2(\mathscr{X}, \mu)$. Let $\{\lambda_k \,|\, k \in \mathbb{I}_+\}$ be the set (at most countable) of all strictly positive eigenvalues of T_μ. We denote by $\widetilde{\varphi}_k \in L^2(\mathscr{X}, \mu)$ their associated eigenfunctions, i.e., in $L^2(\mathscr{X}, \mu)$,

$$\forall k \in \mathbb{I}_+, \quad T_\mu[\widetilde{\varphi}_k] = \lambda_k \widetilde{\varphi}_k, \text{ with } \lambda_k > 0\,,$$

chosen to be orthonormal in $L^2(\mathcal{X}, \mu)$. We also introduce their canonical extensions $\varphi_k \in \mathcal{H}$,

$$\forall x \in \mathscr{X}, \quad \varphi_k(x) = \frac{1}{\lambda_k} T_\mu[\widetilde{\varphi}_k](x)\,, \; k \in \mathbb{I}_+\,, \tag{2.1}$$

so that $\{\sqrt{\lambda_k}\varphi_k \,|\, k \in \mathbb{I}_+\}$ forms an orthonormal basis of $\mathcal{H}_\mu$ for the Hilbert structure of $\mathcal{H}$, see Gauthier and Pronzato (2014, Prop. 3.1).

2.3 Hilbert space decomposition

To the orthogonal decomposition $\mathcal{H} = \mathcal{H}_\mu \overset{\perp}{\oplus} \mathcal{H}_0$ of Section 2.2 corresponds the orthogonal decomposition

$$\mathbb{H} = \mathbb{H}_\mu \overset{\perp}{\oplus} \mathbb{H}_0 \tag{2.2}$$

via the isometry $\mathcal{I}$. We denote by $P_{\mathbb{H}_\mu}$ and $P_{\mathbb{H}_0}$ the orthogonal projections of $\mathbb{H}$ onto $\mathbb{H}_\mu$ and $\mathbb{H}_0$ respectively. The covariance kernel $K_\mu(\cdot,\cdot)$ of the RF $\left(P_{\mathbb{H}_\mu}[Z_x]\right)_{x\in\mathscr{X}}$ is given by

$$\forall x \text{ and } y \in \mathscr{X}, \quad K_\mu(x,y) = \sum_{k\in\mathbb{I}_+} \lambda_k \varphi_k(x)\varphi_k(y),$$

and $K_0(\cdot,\cdot) = K(\cdot,\cdot) - K_\mu(\cdot,\cdot)$ is the covariance of $\left(P_{\mathbb{H}_0}[Z_x]\right)_{x\in\mathscr{X}}$.

For all $k \in \mathbb{I}_+$, we introduce the r.v. $\xi_k = \mathcal{I}\left(\sqrt{\lambda_k}\varphi_k\right)$, so that by construction $\{\xi_k \,|\, k \in \mathbb{I}_+\}$ is an orthonormal basis of $\mathbb{H}_\mu$ (see also Remark 2.1). We then have the following decomposition (or Karhunen-Loève expansion) in $L^2(\Omega,\mathbb{P})$:

$$\forall x \in \mathscr{X}, \quad P_{\mathbb{H}_\mu}[Z_x] = \sum_{k\in\mathbb{I}_+} \sqrt{\lambda_k}\xi_k\varphi_k(x). \tag{2.3}$$

Remark 2.1. If we assume that the realisations of the RF Z belong to $L^2(\mathscr{X},\mu)$ with $\mathbb{P}$-probability one, then the r.v. ξ_k have the following representation: $\xi_k = (1/\sqrt{\lambda_k}) \int_{\mathscr{X}} Z_x \widetilde{\varphi}_k(x)\mathrm{d}\mu(x)$. However, this assumption is stronger than those used in Section 2.2 and is not essential for our study. ◁

2.4 IMSE and truncated-IMSE

Let $\mathbb{H}_D$ be a closed linear subspace of $\mathbb{H}$ and denote by $P_{\mathbb{H}_D}$ the orthogonal projection of $\mathbb{H}$ onto $\mathbb{H}_D$. For $x \in \mathscr{X}$, the r.v. $P_{\mathbb{H}_D}[Z_x]$ is the best linear prediction (unbiased and in terms of the MSE) of the r.v. Z_x relatively to $\mathbb{H}_D$. This prediction is optimal in the Gaussian case and then corresponds to the conditional mean of Z_x relatively to $\mathbb{H}_D$. The IMSE associated with $\mathbb{H}_D$ is given by

$$\mathsf{IMSE}(\mathbb{H}_D) = \int_{\mathscr{X}} \mathsf{E}\left[(Z_x - P_{\mathbb{H}_D}[Z_x])^2\right] \mathrm{d}\mu(x).$$

For $\mathbb{I}_{trc}$ a subset of $\mathbb{I}_+$ (truncation subset), we introduce $\mathbb{H}_{trc} = \overline{\mathrm{span}\,\{\xi_k | k \in \mathbb{I}_{trc}\}}^{L^2(\Omega,\mathbb{P})}$, the closure in $L^2(\Omega,\mathbb{P})$ of the linear space spanned by the r.v. ξ_k, $k \in \mathbb{I}_{trc}$. The truncated-IMSE, with truncation subset $\mathbb{I}_{trc}$, is defined by

$$\mathsf{IMSE}_{trc}(\mathbb{H}_D) = \int_{\mathscr{X}} \mathsf{E}\Big[\big(P_{\mathbb{H}_{trc}}\big[Z_x - P_{\mathbb{H}_D}[Z_x]\big]\big)^2\Big] \mathrm{d}\mu(x). \tag{2.4}$$

More details concerning the truncated-IMSE can be found in (Gauthier and Pronzato, 2014); see also Harari and Steinberg (2014). We have in particular:

$$\mathsf{IMSE}_{trc}(\mathbb{H}_D) \leqslant \mathsf{IMSE}(\mathbb{H}_D) \leqslant \mathsf{IMSE}_{trc}(\mathbb{H}_D) + \sum_{k\notin\mathbb{I}_{trc}} \lambda_k, \tag{2.5}$$

where $k \notin \mathbb{I}_{trc}$ stands for $k \in \mathbb{I}_+ \backslash \mathbb{I}_{trc}$.

An n-point design $D_n = \{x_1, \cdots, x_n\}$ (with $n \in \mathbb{N}^*$ and $x_i \in \mathscr{X}$) is canonically associated with the subspace $\mathbb{H}_{D_n} = \mathrm{span}(Z_{x_1}, \cdots, Z_{x_n})$ of $\mathbb{H}$. We shall use the notation $\mathsf{IMSE}(D_n) = \mathsf{IMSE}(\mathbb{H}_{D_n})$ to refer to the IMSE associated with the design D_n.

3 Spectral truncation and Bayesian linear models

3.1 Interpretation of the random field model as a Bayesian linear model

Consider a (finite) truncation subset $\mathbb{I}_{trc}$ of $\mathbb{I}_+$ (usually corresponding to the n_{trc} largest eigenvalues of T_μ). From (2.2) and (2.3), we have, in $L^2(\Omega, \mathbb{P})$,

$$\forall x \in \mathscr{X}, \quad Z_x = \sum_{k \in \mathbb{I}_{trc}} \beta_k \varphi_k(x) + E_x, \tag{3.1}$$

with $\beta_k = \sqrt{\lambda_k}\xi_k$ and $E_x = \sum_{k \notin \mathbb{I}_{trc}} \sqrt{\lambda_k}\xi_k \varphi_k(x) + P_{\mathbb{H}_0}[Z_x]$. The β_k, $k \in \mathbb{I}_{trc}$, are therefore mutually orthogonal centered r.v. in $L^2(\Omega, \mathbb{P})$ with variance λ_k. Also, $(E_x)_{x \in \mathscr{X}}$ is a centered RF with covariance given by

$$\forall x \text{ and } y \in \mathscr{X}, \quad K_{err}(x, y) = K(x, y) - K_{trc}(x, y), \tag{3.2}$$

where $K_{trc}(x, y) = \sum_{k \in \mathbb{I}_{trc}} \lambda_k \varphi_k(x) \varphi_k(y)$. In addition, for all $x \in \mathscr{X}$, the β_k are orthogonal to E_x.

According to (3.1), we can thus interpret the RF model Z as a Bayesian Linear Model (BLM) with functions φ_k as regressors, $k \in \mathbb{I}_{trc}$, a given prior on the coefficients β_k, and observation errors E_x. We shall refer to (3.1) as the exact BLM induced by Z and the truncation subset $\mathbb{I}_{trc}$.

With vector-matrix notation, we shall denote $\boldsymbol{\beta}$ the (column) random vector with components β_k, and $\boldsymbol{\phi}_{trc}(x)$ the (column) vector with components $\varphi_k(x)$, $k \in \mathbb{I}_{trc}$, $x \in \mathscr{X}$, so that (3.1) becomes

$$Z_x = \boldsymbol{\phi}_{trc}^T(x)\boldsymbol{\beta} + E_x.$$

We shall also denote $\boldsymbol{\Lambda}_{trc} = \mathrm{diag}(\lambda_k | k \in \mathbb{I}_{trc})$ the covariance matrix of the random vector $\boldsymbol{\beta}$; we thus have in particular $K_{trc}(x, y) = \boldsymbol{\phi}_{trc}^T(x)\boldsymbol{\Lambda}_{trc}\boldsymbol{\phi}_{trc}(y)$.

3.2 IMSE for the exact Bayesian linear model

Consider the exact BLM (3.1) for an n-point design $D_n = \{x_1, \cdots, x_n\}$. Define the design matrix $\boldsymbol{\Phi}_{trc}$, with i, k entry $\varphi_k(x_i)$, $1 \leqslant i \leqslant n$ and $k \in \mathbb{I}_{trc}$, and the covariance matrix $\mathbf{K}_{err}$ of the observation errors $(E_{x_1}, \ldots, E_{x_n})$, with i, j entry $K_{err}(x_i, x_j)$, see (3.2). The covariance matrix $\mathbf{K}$ of the vector of observations $\mathbf{z} = (Z_{x_1}, \cdots, Z_{x_n})^T$ is then given by $\mathbf{K} = \boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T + \mathbf{K}_{err}$. For the sake of simplicity, we assume that the design D_n is such that $\mathbf{K}_{err}$ (and thus $\mathbf{K}$) is invertible, but extension to singular matrices is possible through the use of generalised inverses.

We consider the following estimator of $\boldsymbol{\beta}$:

$$\widehat{\boldsymbol{\beta}} = \left(\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\boldsymbol{\Phi}_{trc} + \boldsymbol{\Lambda}_{trc}^{-1}\right)^{-1}\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\mathbf{z},$$

which is solution of the regularised least-squares problem defined by the minimisation of

$$\mathcal{L}^2(\boldsymbol{\beta}) = (\mathbf{z} - \boldsymbol{\Phi}_{trc}\boldsymbol{\beta})^T\mathbf{K}_{err}^{-1}(\mathbf{z} - \boldsymbol{\Phi}_{trc}\boldsymbol{\beta}) + (\boldsymbol{\beta} - \boldsymbol{\beta}_0)^T\boldsymbol{\Lambda}_{trc}^{-1}(\boldsymbol{\beta} - \boldsymbol{\beta}_0), \tag{3.3}$$

with $\boldsymbol{\beta}_0 = \mathsf{E}(\boldsymbol{\beta}) = 0$. Note that when Z is Gaussian, $\widehat{\boldsymbol{\beta}}$ is simply the posterior mean of $\boldsymbol{\beta}$. The Mean-Squared prediction Error (MSE) at $x \in \mathscr{X}$ for the exact BLM is

$$\mathsf{MSE}_{trc}^{K_{err}}(x; D_n) = \mathsf{E}\left\{\left[\boldsymbol{\phi}^T(x)(\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}})\right]^2\right\} = \boldsymbol{\phi}_{trc}^T(x)\left(\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\boldsymbol{\Phi}_{trc} + \boldsymbol{\Lambda}_{trc}^{-1}\right)^{-1}\boldsymbol{\phi}_{trc}(x). \tag{3.4}$$

Remark 3.1. From the Sherman-Morrison-Woodbury matrix identity, we have

$$\begin{aligned}\left(\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\boldsymbol{\Phi}_{trc}+\boldsymbol{\Lambda}_{trc}^{-1}\right)^{-1} &= \boldsymbol{\Lambda}_{trc}-\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T(\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T+\mathbf{K}_{err})^{-1}\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc} \\ &= \boldsymbol{\Lambda}_{trc}-\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T\mathbf{K}^{-1}\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}.\end{aligned} \tag{3.5}$$

For $x\in\mathcal{X}$, the prediction $\boldsymbol{\phi}_{trc}^T(x)\widehat{\boldsymbol{\beta}}$ can be written as

$$\boldsymbol{\phi}_{trc}^T(x)\widehat{\boldsymbol{\beta}}=\boldsymbol{\phi}_{trc}^T(x)\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T(\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T+\mathbf{K}_{err})^{-1}\mathbf{z}.$$

It corresponds to the usual kriging predictor for a (centered) RF model with covariance kernel $K_{trc}(\cdot,\cdot)$ combined with centered observation errors with covariance $K_{err}(\cdot,\cdot)$. In the same way, we obtain for the MSE

$$\mathsf{MSE}_{trc}^{K_{err}}(x;D_n)=K_{trc}(x,x)-\boldsymbol{\phi}_{trc}^T(x)\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T(\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T+\mathbf{K}_{err})^{-1}\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\phi}_{trc}(x),$$

which is, as expected, the usual (simple) kriging variance for a RF model with observation errors. $\triangleleft$

Integrating (3.4) with respect to μ and applying Tonelli's theorem, we obtain

$$\begin{aligned}\mathsf{IMSE}_{trc}^{K_{err}}(D_n) &= \int_{\mathscr{X}}\mathsf{MSE}_{trc}^{K_{err}}(x;D_n)\mathrm{d}\mu(x) \\ &= \operatorname{trace}\left\{(\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\boldsymbol{\Phi}_{trc}+\boldsymbol{\Lambda}_{trc}^{-1})^{-1}\int_{\mathscr{X}}\boldsymbol{\phi}_{trc}(x)\boldsymbol{\phi}_{trc}^T(x)\mathrm{d}\mu(x)\right\} \\ &= \operatorname{trace}\left\{(\boldsymbol{\Phi}_{trc}^T\mathbf{K}_{err}^{-1}\boldsymbol{\Phi}_{trc}+\boldsymbol{\Lambda}_{trc}^{-1})^{-1}\right\},\end{aligned} \tag{3.6}$$

where we have used the property that the eigenfunctions φ_k are orthonormal in $L^2(\mathscr{X},\mu)$. Using (3.5), we obtain that $\mathsf{IMSE}_{trc}^{K_{err}}(D_n)$ given by (3.6) coincides with the truncated-IMSE (2.4) for the design D_n, i.e., $\mathsf{IMSE}_{trc}^{K_{err}}(D_n)=\mathsf{IMSE}_{trc}(D_n)$, which can be written as

$$\mathsf{IMSE}_{trc}(D_n)=\operatorname{trace}\left(\boldsymbol{\Lambda}_{trc}-\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T\mathbf{K}^{-1}\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\right)=\operatorname{trace}\left(\boldsymbol{\Lambda}_{trc}\right)-\operatorname{trace}\left(\mathbf{K}^{-1}\boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}^2\boldsymbol{\Phi}_{trc}^T\right),$$

see Gauthier and Pronzato (2014). Notice that although the model (3.1) is exact, it only gives access to the truncated version of $\mathsf{IMSE}(D_n)$.

3.3 Approximate Bayesian linear model

The main motivation for interpreting a RF model as a BLM is to be in a position to use the classical machinery of approximate-design theory for linear models. However, this cannot be applied directly to the exact model (3.1) where the observation errors are correlated. We therefore introduce an approximate linear model with uncorrelated errors and consider

$$\forall x\in\mathscr{X},\quad \widetilde{Z}_x=\sum_{k\in\mathbb{I}_{trc}}\beta_k\varphi_k(x)+\mathcal{E}_x, \tag{3.7}$$

where the β_k and φ_k are defined as in (3.1), and where the errors $\mathcal{E}_x\in L^2(\Omega,\mathbb{P})$ are centered, orthogonal to the β_k, $k\in\mathbb{I}_{trc}$, and such that, for x and $y\in\mathscr{X}$,

$$\mathsf{E}\left(\mathcal{E}_x\mathcal{E}_y\right)=\Sigma(x,y)=\begin{cases}\sigma^2(x) & \text{if } x=y,\\ 0 & \text{otherwise.}\end{cases}$$

In order to ensure that the two models (3.1) and (3.7) have the same integrated variance with respect to μ, we impose that the variance $\sigma^2(x)$ satisfies

$$\tau_{err} = \int_{\mathscr{X}} K_{err}(x,x)\mathrm{d}\mu(x) = \sum_{k \notin \mathbb{I}_{trc}} \lambda_k = \tau - \tau_{trc} = \int_{\mathscr{X}} \sigma^2(x)\mathrm{d}\mu(x), \tag{3.8}$$

with $\tau_{trc} = \sum_{k\in\mathbb{I}_{trc}} \lambda_k$.

For an n-point design $D_n = \{x_1, \cdots, x_n\}$, we denote by $\boldsymbol{\Sigma}$ the (diagonal) covariance matrix of the observation errors, with i,j entry $\Sigma(x_i,x_j)$. The covariance matrix $\widetilde{\mathbf{K}}$ of the vector of observations $\widetilde{\mathbf{z}} = (\widetilde{Z}_{x_1}, \cdots, \widetilde{Z}_{x_n})^T$ is thus $\widetilde{\mathbf{K}} = \boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\Phi}_{trc}^T + \boldsymbol{\Sigma}$. As in Section 3.2, we assume for the sake of simplicity that $\boldsymbol{\Sigma}$ (and thus $\widetilde{\mathbf{K}}$) is invertible. The IMSE calculated with the approximate model (3.7) is then given by

$$\mathsf{IMSE}^{\Sigma}_{trc}(D_n) = \mathrm{trace}\left[(\boldsymbol{\Phi}_{trc}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}_{trc} + \boldsymbol{\Lambda}_{trc}^{-1})^{-1}\right].$$

Homoscedastic errors with $\sigma^2(x) = \sigma^2$ constant are considered in (Spöck and Pilz, 2010), which gives (assuming the measure μ is finite) $\sigma^2 = \tau_{err}/\mu(\mathscr{X})$ for any $x \in \mathscr{X}$. However, even in the case when $K(x,x)$ is constant, the function $x \in \mathscr{X} \mapsto K_{err}(x,x)$ is generally strongly oscillating due to the form of the eigenfunctions that enter $K_{trc}(x,x)$, see Figs. 1 and 2 of Section 7.1. Choosing an heteroscedastic model with $\sigma^2(x) = K_{err}(x,x)$ seems therefore more appropriate than $\sigma^2(x) = \sigma^2$. Note that it amounts to replacing $\mathbf{K}_{err}$ by its diagonal in the developments of Section 3.2 and gives $\mathsf{Var}(\widetilde{Z}_x) = \mathsf{Var}(Z_x)$.

4 Truncated-IMSE and continuous-design relaxation

We consider a linear model of the form (3.7) with orthogonal observation errors having variance $\sigma^2(\cdot)$. We assume that $x \mapsto \sigma^2(x)$ is measurable on $(\mathscr{X},\mathcal{A})$, with $\sigma^2(x) > \varepsilon > 0$ for all $x \in \mathscr{X}$ (see Remark 4.1), and that $K(x,x)$ is bounded on $\mathscr{X}$ by some constant C. This implies that

$$\forall x \in \mathscr{X}\,, \quad |\varphi_k(x)| = |(\varphi_k|K_x)_{\mathcal{H}}| \leq \|\varphi_k\|_{\mathcal{H}}\,\|K_x\|_{\mathcal{H}} = \sqrt{K(x,x)/\lambda_k} \leq \sqrt{C/\lambda_k}\,,$$

and $\varphi_k(x)$ is bounded on $\mathscr{X}$. These assumptions ensure that the information matrix

$$\mathbf{M}_{\nu} = \int_{\mathscr{X}} \frac{1}{\sigma^2(x)}\boldsymbol{\phi}_{trc}(x)\boldsymbol{\phi}_{trc}^T(x)\,\mathrm{d}\nu(x), \tag{4.1}$$

with k,l entry $\int_{\mathscr{X}}[1/\sigma^2(x)]\varphi_k(x)\varphi_l(x)\mathrm{d}\,\nu(x)$, is well defined for any measure ν in the set $\mathscr{P}_{\mathcal{A}}$ of probability measures on $(\mathscr{X},\mathcal{A})$: $\mathbf{M}_{\nu}$ is bounded, symmetric and non-negative definite.

When ν is the empirical measure $\nu_n = (1/n)\sum_{i=1}^n \delta_{x_i}$ associated with D_n, where δ_{x_i} stands for the Dirac measure at x_i, we obtain $\mathbf{M}_{\nu_n} = (1/n)\,\boldsymbol{\Phi}_{trc}^T\boldsymbol{\Sigma}^{-1}\boldsymbol{\Phi}_{trc}$. Although such empirical measures are the only ones that can be implemented in the form of an exact design without replication, we shall consider relaxed optimisation problems involving general measures ν. Once a measure ν^* optimal (or close enough to optimality) will be determined, its support will be used in Section 5.2 to generate an n-point design D_n.

For any fixed $\alpha \in \mathbb{R}^+$, we define the following functional $\Psi_\alpha(\cdot)$ on $\mathscr{P}_{\mathcal{A}}$,

$$\forall \nu \in \mathscr{P}_{\mathcal{A}}, \quad \Psi_\alpha(\nu) = \mathrm{trace}\left[(\alpha\mathbf{M}_{\nu} + \boldsymbol{\Lambda}_{trc}^{-1})^{-1}\right], \tag{4.2}$$

so that $\Psi_n(\nu_n) = \mathsf{IMSE}^{\Sigma}_{trc}(D_n)$ for the empirical measure ν_n. The parameter α has a strong impact on the optimal measure that minimises $\Psi_\alpha(\cdot)$, as illustrated in Section 7. One can readily check that, for a fixed $\nu \in \mathscr{P}_{\mathcal{A}}$, the function $\alpha \in \mathbb{R}^+ \mapsto \Psi_\alpha(\nu)$ is positive, decreasing, and satisfies

$$\Psi_0(\nu) = \sum_{k \in \mathbb{I}_{trc}} \lambda_k \text{ and } \lim_{\alpha \to +\infty} \Psi_\alpha(\nu) = 0.$$

For example, for the particular measure $\mathrm{d}\check{\nu}(x) = [\sigma^2(x)/\tau_{err}]\,\mathrm{d}\mu(x)$, with $\tau_{err} = \tau - \tau_{trc}$ and $\sigma^2(\cdot)$ satisfying (3.8), we obtain $\mathbf{M}_{\check{\nu}} = \mathrm{Id}/\tau_{err}$ and $\Psi_\alpha(\check{\nu}) = \sum_{k \in \mathbb{I}_{trc}} (\alpha/\tau_{err} + 1/\lambda_k)^{-1}$.

The functional $\Psi_\alpha(\cdot)$ defined by (4.2) corresponds to a Bayesian A-optimality criterion, see Pilz (1983); Chaloner (1984); Pukelsheim (1993, Chap. 11), which is convex in $\nu \in \mathscr{P}_{\mathcal{A}}$ and non-increasing for Loewner ordering. In order to be able to address the minimisation of $\Psi_\alpha(\nu)$ with respect to $\nu \in \mathscr{P}_{\mathcal{A}}$, we assume additionally that the set $\{\boldsymbol{\phi}_{trc}(x)\boldsymbol{\phi}^T_{trc}(x)/\sigma^2(x) | x \in \mathscr{X}\}$ is compact. This is not too restrictive, and is satisfied in particular when $\mathscr{X}$ is finite, or is a compact subset of $\mathbb{R}^d$ with $\sigma(\cdot)$ and $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ being continuous (so that the $\varphi_k(\cdot)$ are continuous on $\mathscr{X}$), see the examples in Section 7.

Then, from classical results in optimum experimental design (approximate theory), there exists at least one probability measure $\nu^* \in \mathscr{P}_{\mathcal{A}}$ which minimises $\Psi_\alpha(\cdot)$ and has $m \leqslant 1 + n_{trc}(n_{trc}+1)/2$ support points in $\mathscr{X}$ (see, e.g., Winkler (1988), Barvinok (2002, Chap. 3)); moreover, the optimal matrix $\mathbf{M}_{\nu^*}$ is unique. The directional derivative of $\Psi_\alpha(\cdot)$ at ν in the direction $\eta - \nu$ is given by

$$\begin{aligned} F_{\Psi_\alpha}(\nu, \eta) &= \lim_{\gamma \to 0^+} \frac{\Psi_\alpha[(1-\gamma)\nu + \gamma\eta] - \Psi_\alpha(\nu)}{\gamma} \\ &= -\alpha \operatorname{trace}\left\{(\alpha\mathbf{M}_\nu + \boldsymbol{\Lambda}^{-1}_{trc})^{-1}(\mathbf{M}_\eta - \mathbf{M}_\nu)(\alpha\mathbf{M}_\nu + \boldsymbol{\Lambda}^{-1}_{trc})^{-1}\right\}, \end{aligned} \tag{4.3}$$

and the convexity and differentiability of $\Psi_\alpha(\cdot)$ imply that a measure $\nu^* \in \mathcal{P}$ is optimal if and only if $F_{\Psi_\alpha}(\nu^*, \eta) \geqslant 0$ for any $\eta \in \mathcal{P}$. This corresponds to the Equivalence Theorem for Bayesian A-optimal design stated below, see Pilz (1983).

Theorem 4.1. *For any given $\alpha > 0$, the measure $\nu^* \in \mathcal{P}_{\mathcal{A}}$ minimises $\Psi_\alpha(\cdot)$ if and only if, for all $x \in \mathscr{X}$,*

$$\boldsymbol{\phi}^T_{trc}(x)(\alpha\mathbf{M}_{\nu^*} + \boldsymbol{\Lambda}^{-1}_{trc})^{-2}\boldsymbol{\phi}_{trc}(x) \leqslant \sigma^2(x) \operatorname{trace}\left\{\mathbf{M}_{\nu^*}(\alpha\mathbf{M}_{\nu^*} + \boldsymbol{\Lambda}^{-1}_{trc})^{-2}\right\}. \tag{4.4}$$

Equality is achieved in (4.4) ν^*-almost everywhere (that is, for any support point of ν^*). The convexity and differentiability of $\Psi_\alpha(\cdot)$ imply that, for any $\nu \in \mathcal{P}_{\mathcal{A}}$,

$$\Psi_\alpha(\nu) \leqslant \Psi_\alpha(\nu^*) - \min_{x \in \mathscr{X}} F_{\Psi_\alpha}(\nu, \delta_x), \tag{4.5}$$

which can be used to check distance from optimality. Many efficient convex optimisation algorithms are available for the construction of a sequence of measures $\nu^{(k)}$ such that $\Psi_\alpha(\nu^{(k)})$ converges to $\Psi_\alpha(\nu^*)$, see, e.g., Pronzato and Pázman (2013, Chap. 9).

Remark 4.1. Assume that there exists $x_0 \in \mathscr{X}$ such that $\sigma^2(x_0) = 0$; we distinguish two cases.

1. There exists $k \in \mathbb{I}_{trc}$ such that $\varphi_k(x_0) \neq 0$ (and therefore, $K_{trc}(x_0, x_0) \neq 0$). One should then include x_0 in the design since it allows observation of the exact value of $\boldsymbol{\phi}^T_{trc}(x_0)\boldsymbol{\beta}$.

2. $K_{trc}(x_0, x_0) = 0$. In that case, we can roughly say that there is nothing to learn in x_0, and we may exclude x_0 from the search space. This can be achieved for instance by considering the pseudo-inverse $(\sigma^2(x))^\dagger$ of $\sigma^2(x)$ in (4.1), with $r^\dagger = 1/r$ if $r \neq 0$, and $r^\dagger = 0$ if $r = 0$. $\triangleleft$

5 Numerical approach

In this section we discuss the implementation of the method proposed in Sections 3 and 4, assuming that a pointwise quadrature is used to approximate the integral of the MSE and restricting the design space to quadrature points. A similar approach could be used for any design region formed by a finite set of points. We also consider the extraction of an n-point design from an optimal measure for $\Psi_\alpha(\cdot)$. The methods presented in this section can be straightforwardly applied to the framework of Section 6 involving RF models that include an unknown linear parametric trend.

5.1 Quadrature approximation and quadrature-restricted continuous design

Assume that the measure μ has the following form

$$\mu = \sum_{j=1}^{N_q} \omega_j \delta_{s_j}, \tag{5.1}$$

with N_q quadrature points $s_j \in \mathscr{X}$ receiving weights $\omega_j > 0$. We introduce the two $N_q \times N_q$ matrices $\mathbf{W} = \operatorname{diag}\left(\omega_1, \cdots, \omega_{N_q}\right)$ and $\mathbf{Q}$ with i,j term $\mathbf{Q}_{i,j} = K(s_i, s_j)$, for $1 \leqslant i, j \leqslant N_q$; $\mathbf{W}$ is thus the diagonal matrix of quadrature weights and $\mathbf{Q}$ is the covariance matrix for quadrature points.

Consider the spectral decomposition of the matrix $\mathbf{QW}$ in the Hilbert space $\mathbb{R}^{N_q}$ endowed with the inner product $(\cdot|\cdot)_{\mathbf{W}}$, with, for $\mathbf{x}$ and $\mathbf{y} \in \mathbb{R}^{N_q}$, $(\mathbf{x}|\mathbf{y})_{\mathbf{W}} = \mathbf{x}^T\mathbf{W}\mathbf{y}$, see Gauthier and Pronzato (2014, 2015a) for more details. We denote by $\lambda_1 \geqslant \lambda_2 \geqslant \cdots \geqslant \lambda_{N_q} \geqslant 0$ the eigenvalues of the matrix $\mathbf{QW}$ and by $\mathbf{v}_1, \ldots, \mathbf{v}_{N_q}$ the corresponding eigenvectors, i.e., $\mathbf{QW} = \mathbf{P}\mathbf{\Lambda}\mathbf{P}^{-1}$ with $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \ldots, \lambda_{N_q})$ and $\mathbf{P} = (\mathbf{v}_1|\cdots|\mathbf{v}_{N_q})$. Then, $\left\{\mathbf{v}_1, \ldots, \mathbf{v}_{N_q}\right\}$ forms an orthonormal basis of $\mathbb{R}^{N_q}$ for the inner product $(\cdot|\cdot)_{\mathbf{W}}$, so that $\mathbf{P}^T\mathbf{W}\mathbf{P} = \mathrm{Id}_{N_q}$, the N_q-dimensional identity matrix.

Let $\mathscr{P}_{\mathcal{A},\mu} \subset \mathscr{P}_{\mathcal{A}}$ denote the subset of probability measures dominated by μ given by (5.1), i.e., of all measures $\nu = \sum_{j=1}^{N_q} p_j \delta_{s_j}$, with $p_j \geqslant 0$ and $\sum_{j=1}^{N_q} p_j = 1$. For a truncation set $\mathbb{I}_{trc}$, the matrix $\mathbf{M}_\nu$ given by (4.1) and associated with a measure $\nu \in \mathscr{P}_{\mathcal{A},\mu}$ is then given by

$$\mathbf{M}_\nu = (\mathbf{P}_{\cdot,\mathbb{I}_{trc}})^T \mathbf{\Omega} (\mathbf{P}_{\cdot,\mathbb{I}_{trc}}), \tag{5.2}$$

where $\mathbf{\Omega} = \operatorname{diag}\left(p_1/\sigma^2(s_1), \cdots, p_{N_q}/\sigma^2(s_{N_q})\right)$ and $\mathbf{P}_{\cdot,\mathbb{I}_{trc}}$ stands for the $N_q \times n_{trc}$ matrix consisting of the columns of $\mathbf{P}$ with index in $\mathbb{I}_{trc}$.

5.2 Design extraction

Suppose that an optimal design measure ν_m^* minimizing $\Psi_\alpha(\nu)$ given by (4.2) and (5.2) has been determined, ν_m^* having support $\mathcal{S}_m^* = \{s_{i_1}, \cdots, s_{i_m}\}$. We must still define a procedure for extracting an n-point design D_n from ν_m^*, for a given n. Note that the issue differs from the usual rounding problem in approximate design theory, see, e.g., Fedorov (1972, p. 157); Pukelsheim and Reider (1992), since we request exactly one observation per point and, moreover, the scalar α and number n_{trc} of regressors in the underlying linear model can be used as tuning parameters.

First, note that the values of n and n_{trc} cannot be chosen in total independence: indeed, for a given n, it seems in general illusory to predict the response of a model involving much more than n regressors. Taking $n_{trc} \approx n$ thus seems reasonable. Second, since $\Psi_n(\nu_n) = \mathsf{IMSE}_{trc}^{\Sigma}(D_n)$ for the empirical measure ν_n, see Section 4, choosing $\alpha \approx n$ seems reasonable too. Optimal design measures

ν_m^* for values of α and n_{trc} close to the requested n tend to be constituted of n' points, or n' clusters of neighboring points, with $n' \approx n$, all points or clusters of points receiving similar weights. One may then proceed by trial an error, until $n' = n$ or is close enough to n, and merge points within eventual clusters.

Various approaches can be used to automatize this construction, e.g., aggregation of support points via minimum-spanning-tree clustering, branch-and-bound-type methods, etc. In what follows we describe a greedy merging strategy (Algorithm 1) which is rather straightforward to implement, has low complexity (of order $\mathcal{O}(m^3)$) and proved rather efficient on the numerical tests that we performed. A few examples are given in Section 7. If $\nu = \sum_{j=1}^{n_s} p_j \delta_{s_{i_j}}$ is the current design measure, supported by n_s quadrature points s_{i_j} with $i_j \in \{1, \cdots, N_q\}$ and $p_j > 0$ for all j with $\sum_{j=1}^{n_s} p_j = 1$, the algorithm considers all $n_s(n_s-1)/2$ possible measures $\nu_{[a\to b]}$ obtained by transferring the weight p_a from s_{i_a} to s_{i_b} ($\nu_{[a\to b]}$ is thus supported by $n_s - 1$ points, s_{i_a} being removed from ν). The measure $\nu_{[a^*\to b^*]}$ with smallest value of $\Psi_\alpha(\cdot)$ is then carried forward to the next iteration. The current measure ν after k iterations is such that $\Psi_\alpha(\nu) = \psi_k$ and has $m - k$ support points, the sequence $I_1, \ldots, I_k$ contains the indices of the k quadrature points that have been removed from the support $\mathcal{S}_m^*$ of ν_m^*.

Algorithm 1 Greedy algorithm for design extraction

Require: $\nu_m^* \in \mathscr{P}_{\mathcal{A},\mu}$ (with m support-points $\mathcal{S}_m^*$) and $\Psi_\alpha(\cdot)$;
1: $\nu \leftarrow \nu_m^*$; $n_s \leftarrow m$; $\psi_0 \leftarrow \Psi_\alpha(\nu_m^*)$;
2: **while** $n_s > 1$ **do**
3: search for $\{a^*, b^*\} = \operatorname{argmin} \Psi_\alpha(\nu_{[a\to b]})$, with $a \in \{1, \cdots, n_s\}$ and $b \in \{1, \cdots, n_s\} \backslash \{a\}$;
4: $n_s \leftarrow n_s - 1$; $\psi_{m-n_s} \leftarrow \Psi_\alpha(\nu_{[a^*\to b^*]})$; $\nu \leftarrow \nu_{[a^*\to b^*]}$; $I_{m-n_s} \leftarrow i_{a^*}$;
5: **end while**
6: **return** $\{\psi_k\}$ (Ψ_α values) and $\{I_k\}$ (points removed from $\mathcal{S}_m^*$).

The sequence $\{\psi_k\}$ is generally non-decreasing (due to the fact that ν_m^* is optimal for $\Psi_\alpha(\cdot)$), and we can stop removing points when ψ_k is significantly larger than $\psi_0 = \Psi_\alpha(\nu_m^*)$. Usually, this occurs for an abrupt increase of Ψ_α, and the size n of the design extracted is then chosen equal to the number of support points just before the jump. We shall denote by D_n^{supp} this design, and by D_n^{ext} the best n-point design obtained by running a local descent algorithm, restricted to quadrature points and initialized at D_n^{supp}, see Gauthier and Pronzato (2015a). Section 7 will present some examples.

Remark 5.1. A method based on the aggregation of support points of ν_m^* via minimum-spanning-tree clustering has also been considered, with the metric $\Delta(x, x') = K(x, x) + K(x', x') - 2K(x, x')$ induced by K (the design points are given by the weighed barycentres of the clusters, with weights given by the p_k), see Gauthier and Pronzato (2015b). For the examples we have treated, the designs obtained looked much similar to D_n^{ext} obtained with Algorithm 1. Notice that this approach requires the computation of canonical extensions (2.1), i.e., of summations over the N_q quadrature points, since in general the measure obtained does not belong to $\mathscr{P}_{\mathcal{A},\mu}$.

When $\mathscr{X}$ is a compact subset of $\mathbb{R}^d$, a continuous local minimisation of the IMSE with respect to $D_n \in \mathscr{X}^n$ (i.e., with respect to $n \times d$ variables), initialised at D_n^{ext}, can be performed at reasonable computational cost, using a standard optimisation algorithm (the canonical extensions (2.1) must be used here too). For most covariance kernels the optimal design points lie in the convex hull of

$\mathscr{X}$, and no constraint need to be taken into account if $\mathscr{X}$ is convex. However, the decrease of IMSE compared to D_n^{ext} is usually marginal when the set of quadrature points is dense enough, and we shall not consider this procedure any further in the paper. $\triangleleft$

6 Random field models with unknown linear parametric trend

The approach described in previous sections can be extended to RF models that include an unknown linear parametric trend, as described in (Spöck and Pilz, 2010). Their method is detailed in Sections 6.1 and 6.2. When prior information on the trend parameters is available, we show in Section 6.3 how a rather straightforward kernel augmentation allows us to recast the problem as one involving a RF with known trend. In absence of informative prior on the trend parameters, a kernel reduction is proposed in Section 6.4 that permits to avoid confusion between the trend and RF behaviours.

6.1 IMSE for a random field with unknown linear parametric trend

We still consider the framework and notation of Section 2, but now the RF is $(Y_x)_{x\in\mathscr{X}}$ such that, for all $x \in \mathscr{X}$,

$$Y_x = \mathbf{g}^T(x)\boldsymbol{\theta} + Z_x, \tag{6.1}$$

where $\mathbf{g}(x) = \big(g_1(x), \cdots, g_p(x)\big)^T$ is a column vector of (known) real-valued functions on $\mathscr{X}$ and where $\boldsymbol{\theta} = \big(\theta_1, \ldots, \theta_p\big)^T \in \mathbb{R}^p$ is an unknown vector, $p \in \mathbb{N}^*$. We denote by $\mathcal{T}$ the linear subspace spanned by the trend-functions $\{g_1, \cdots, g_p\}$. We suppose that $g_j \in L^2(\mathscr{X}, \mu)$ for all $j \in \{1, \cdots, p\}$ and, in view of Section 4, we also assume that the set $\{\mathbf{g}(x)\mathbf{g}^T(x)/\sigma^2(x) | x \in \mathscr{X}\}$ is compact.

Following Spöck and Pilz (2010), we consider a prior on the first two moments of $\boldsymbol{\theta}$ and assume that $\theta_i \in L^2(\Omega, \mathbb{P})$ for all $i \in \{1, \cdots, p\}$ with

$$\mathsf{E}(\boldsymbol{\theta}) = \boldsymbol{\theta}_0 \text{ and } \mathsf{Cov}(\boldsymbol{\theta}) = \mathbf{A}, \tag{6.2}$$

where $\boldsymbol{\theta}_0 \in \mathbb{R}^p$ and $\mathbf{A}$ is a $p \times p$ positive-definite matrix. We also assume that the θ_i are orthogonal to $\mathbb{H}$. The case where no prior information on $\boldsymbol{\theta}$ is available corresponds to replacing $\mathbf{A}^{-1}$ by the null matrix and will be considered into details in Section 6.4.

For the RF model (6.1) and design $D_n = \{x_1, \cdots, x_n\}$, the MSE at $x \in \mathscr{X}$ is given by (assuming that all matrix inverses are well-defined)

$$\begin{aligned}\mathsf{MSE}(x; D_n) = K(x,x) - \mathbf{k}^T(x)\mathbf{K}^{-1}\mathbf{k}(x) \\ + \big[\mathbf{g}(x) - \mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}(x)\big]^T\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\big[\mathbf{g}(x) - \mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}(x)\big],\end{aligned} \tag{6.3}$$

where $\mathbf{G}$ is the $n \times p$ design-matrix with i, j entry $\mathbf{G}_{i,j} = g_j(x_i)$, and where

$$\mathbf{k}(x) = \big(K_{x_1}(x), \cdots, K_{x_n}(x)\big)^T.$$

This is the usual expression obtained in Bayesian kriging, see Omre and Halvorsen (1989), Santner et al. (2003, Chap. 4). The IMSE criterion is then given by

$$\mathsf{IMSE}(D_n) = \int_{\mathscr{X}} \mathsf{MSE}(x; D_n)\mathrm{d}\mu(x). \tag{6.4}$$

6.2 Direct spectral approximation and induced exact BLM

Consider the decomposition (3.1) for a truncation subset $\mathbb{I}_{trc}$. It yields the following exact model (with equality in $L^2(\Omega, \mathbb{P})$),

$$Y_x = \mathbf{f}^T(x)\boldsymbol{\gamma} + E_x, \tag{6.5}$$

where

$$\mathbf{f}(x) = \begin{pmatrix} \mathbf{g}(x) \\ \boldsymbol{\phi}_{trc}(x) \end{pmatrix} \text{ and } \boldsymbol{\gamma} = \begin{pmatrix} \boldsymbol{\theta} \\ \boldsymbol{\beta} \end{pmatrix}, \text{ with } \mathsf{E}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}_0 = \begin{pmatrix} \boldsymbol{\theta}_0 \\ 0 \end{pmatrix} \text{ and } \mathsf{Cov}(\boldsymbol{\gamma}) = \boldsymbol{\Gamma} = \begin{pmatrix} \mathbf{A} & 0 \\ 0 & \boldsymbol{\Lambda}_{trc} \end{pmatrix}. \tag{6.6}$$

For a design $D_n = \{x_1, \cdots, x_n\}$, denote by $\mathbf{y}$ the vector of observations $(Y_{x_1}, \cdots, Y_{x_n})^T$ and let $\mathbf{F} = (\mathbf{G}, \boldsymbol{\Phi}_{trc})$. The underlying regularised least-squares problem yields the estimator

$$\widehat{\boldsymbol{\gamma}} = \big(\mathbf{F}^T\mathbf{K}_{err}^{-1}\mathbf{F} + \boldsymbol{\Gamma}^{-1}\big)^{-1}\big(\mathbf{F}^T\mathbf{K}_{err}^{-1}\mathbf{y} + \boldsymbol{\Gamma}^{-1}\boldsymbol{\gamma}_0\big).$$

The MSE at $x \in \mathscr{X}$ is then

$$\mathsf{MSE}_{trc}^{K_{err}}(x; D_n) = \mathsf{E}\Big(\big[\mathbf{f}^T(x)(\boldsymbol{\gamma} - \widehat{\boldsymbol{\gamma}})\big]^2\Big) = \mathbf{f}^T(x)\big(\mathbf{F}^T\mathbf{K}_{err}^{-1}\mathbf{F} + \boldsymbol{\Gamma}^{-1}\big)^{-1}\mathbf{f}(x). \tag{6.7}$$

By expanding (6.7), we also obtain

$$\begin{aligned}\mathsf{MSE}_{trc}^{K_{err}}(x; D_n) &= K_{trc}(x,x) - \mathbf{k}_{trc}^T(x)\mathbf{K}^{-1}\mathbf{k}_{trc}(x) \\ &\quad + \big[\mathbf{g}(x) - \mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}_{trc}(x)\big]^T\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\big[\mathbf{g}(x) - \mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}_{trc}(x)\big],\end{aligned} \tag{6.8}$$

where we have denoted $\mathbf{k}_{trc}(x) = \boldsymbol{\Phi}_{trc}\boldsymbol{\Lambda}_{trc}\boldsymbol{\phi}_{trc}(x)$. Similarly to Section 3.2, $\mathsf{MSE}_{trc}^{K_{err}}(x; D_n)$ corresponds to the MSE obtained from (6.3) through spectral truncation. Substituting a diagonal matrix for $\mathbf{K}_{err}$ we get an approximated IMSE that can be used for convex design optimisation, see Sections 3.3 and 4.

Denote by $\mathbf{M}_{\mathbf{g}}$ the Gram matrix of the trend functions $g_1, \cdots, g_p$ in $L^2(\mathscr{X}, \mu)$, that is, in matrix notation,

$$\mathbf{M}_{\mathbf{g}} = \int_{\mathscr{X}} \mathbf{g}(x)\mathbf{g}^T(x)\mathrm{d}\mu(x).$$

We assume that $\mathbf{M}_{\mathbf{g}}$ is invertible. The truncated IMSE is obtained by integrating (6.7) with respect to μ,

$$\mathsf{IMSE}_{trc}^{K_{err}}(D_n) = \mathrm{trace}\,\big[\big(\mathbf{F}^T\mathbf{K}_{err}^{-1}\mathbf{F} + \boldsymbol{\Gamma}^{-1}\big)^{-1}\mathbf{U}\big], \tag{6.9}$$

where $\mathbf{U}$ is the Gram matrix of $\mathbf{f}(\cdot)$ in $L^2(\mathscr{X}, \mu)$, i.e.,

$$\mathbf{U} = \int_{\mathscr{X}} \mathbf{f}(x)\mathbf{f}^T(x)\mathrm{d}\mu(x) = \begin{pmatrix} \mathbf{M}_{\mathbf{g}} & \big(\mathbf{g}\big|\boldsymbol{\phi}_{trc}^T\big)_{L^2} \\ \big(\boldsymbol{\phi}_{trc}\big|\mathbf{g}^T\big)_{L^2} & \mathrm{Id}_{n_{trc}} \end{pmatrix}. \tag{6.10}$$

It is instructive to compare the IMSE (6.4) with the truncated-IMSE (6.9). For the initial RF model (6.1), after recombination we obtain $\mathsf{IMSE}(D_n) = \tau + B - C - 2D$, where we have set

$$\begin{aligned} B &= \int_{\mathscr{X}} \mathbf{g}^T(x)\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{g}(x)\mathrm{d}\mu(x), \\ C &= \int_{\mathscr{X}} \mathbf{k}^T(x)\big(\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{G}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{K}^{-1}\big)\mathbf{k}(x)\mathrm{d}\mu(x), \text{ and} \\ D &= \int_{\mathscr{X}} \mathbf{g}^T(x)\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}(x)\mathrm{d}\mu(x). \end{aligned}$$

In the same way, considering (6.8), we get $\mathsf{IMSE}_{trc}^{K_{err}}(D_n) = \tau_{trc} + B - C_{trc} - 2D_{trc}$, with

$$C_{trc} = \int_{\mathscr{X}} \mathbf{k}_{trc}^T(x)\big(\mathbf{K}^{-1} - \mathbf{K}^{-1}\mathbf{G}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{K}^{-1}\big)\mathbf{k}_{trc}(x)\mathrm{d}\mu(x), \text{ and}$$
$$D_{trc} = \int_{\mathscr{X}} \mathbf{g}^T(x)\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{K}^{-1}\mathbf{k}_{trc}(x)\mathrm{d}\mu(x).$$

We then obtain the following (the proof is given in Appendix A).

Proposition 6.1. *For any truncation subset $\mathbb{I}_{trc}$, we have $0 \leqslant C - C_{trc} \leqslant \tau_{err} = \tau - \sum_{k\in\mathbb{I}_{trc}} \lambda_k$.*

On the other hand, one may note that D and D_{trc} respectively involve terms of the form $\int_{\mathscr{X}} K(x,x_i)g_j(x)\mathrm{d}\mu(x) = T_\mu\big[g_j\big](x_i) = \sum_{k\in\mathbb{I}_+} \lambda_k\big(\varphi_k\big|g_j\big)_{L^2}\varphi_k(x_i)$ and $\int_{\mathscr{X}} K_{trc}(x,x_i)g_j(x)\mathrm{d}\mu(x) = \sum_{k\in\mathbb{I}_{trc}} \lambda_k\big(\varphi_k\big|g_j\big)_{L^2}\varphi_k(x_i)$, for $i \in \{1,\cdots,n\}$ and $j \in \{1,\cdots,p\}$. The derivation of general error bounds on the difference $D - D_{trc}$ seems therefore more complicated, which partly motivates the two following sections (see also Section 7.4 for examples).

6.3 Presence of a prior: equivalence with a model with known trend

In presence of an informative prior on $\boldsymbol{\theta}$, the model defined by (6.1) and (6.2) can be interpreted as a RF model with known trend. Indeed, for and x and y in $\mathscr{X}$ we have

$$\mathsf{E}(Y_x) = \mathbf{g}^T(x)\boldsymbol{\theta}_0 \quad \text{and} \quad \mathsf{Cov}(Y_x, Y_y) = \mathbf{g}^T(x)\mathbf{A}\mathbf{g}^T(y) + K(x,y) = K^{\text{full}}(x,y). \tag{6.11}$$

We shall refer to the kernel $K^{\text{full}}(\cdot,\cdot)$ as the *augmented kernel*. Applying the Sherman-Morrison-Woodbury identity (assuming, for the sake of simplicity, that D_n is such that $\mathbf{K} + \mathbf{G}\mathbf{A}\mathbf{G}^T$ is invertible), we can easily check that the two RF models (6.1) and (6.11) yield the same predictions.

We can then consider the integral operator

$$T_\mu^{\text{full}}[f](x) = \int_{\mathscr{X}} f(t)K^{\text{full}}(x,t)\mathrm{d}\mu(x),$$

with $f \in L^2(\mathscr{X},\mu)$ and $x \in \mathscr{X}$, and apply the same approach as in Sections 3 and 4 for models without trend. In particular, bounds similar to (2.5) are available and straightforward calculation shows that

$$\tau^{\text{full}} = \int_{\mathscr{X}} K^{\text{full}}(x,x)\mathrm{d}\mu(x) = \tau + \text{trace}\left(\mathbf{A}\mathbf{M}_{\mathbf{g}}\right) \geqslant \tau.$$

6.4 Absence of informative prior: IMSE-adapted kernel reduction

Here we can take advantage of the non-uniqueness of the kernels associated with a given semi-Hilbert structure. Denote by $\mathfrak{p}$ the orthogonal projection of $L^2(\mathscr{X},\mu)$ onto $\mathcal{T}$ and define $\mathfrak{q} = \mathrm{id}_{L^2} - \mathfrak{p}$. For $f \in L^2(\mathscr{X},\mu)$, we obtain, in matrix notation,

$$\mathfrak{p}f = \mathbf{g}^T\mathbf{M}_{\mathbf{g}}^{-1}\big(\mathbf{g}\big|f\big)_{L^2}.$$

Assume, for the sake of simplicity, that the realisations of $(Z_x)_{x\in\mathscr{X}}$ belongs to $L^2(\mathscr{X},\mu)$ with $\mathbb{P}$-probability 1 (this assumption is not necessary, however, for the construction of the kernel $K^{\mathfrak{q}}(\cdot,\cdot)$ given in (6.13) and for Theorem 6.1). For $x \in \mathscr{X}$, we can then define

$$\mathfrak{p}Z_x = \mathbf{g}^T(x)\mathbf{M}_{\mathbf{g}}^{-1}\int_{\mathscr{X}} \mathbf{g}(t)Z_t\mathrm{d}\mu(t),$$

so that $\mathfrak{p}Z_x \in L^2(\Omega, \mathbb{P})$, $\mathsf{E}(\mathfrak{p}Z_x) = 0$ and, for $y \in \mathscr{X}$,

$$\mathsf{E}\big((\mathfrak{p}Z_x)(\mathfrak{p}Z_y)\big) = \mathbf{g}^T(x)\mathbf{M}_{\mathbf{g}}^{-1}\big(T_\mu[\mathbf{g}]\big|\mathbf{g}^T\big)_{L^2}\mathbf{M}_{\mathbf{g}}^{-1}\mathbf{g}^T(x) \text{ and } \mathsf{E}\big((\mathfrak{p}Z_x)Z_y\big) = \mathbf{g}^T(x)\mathbf{M}_{\mathbf{g}}^{-1}T_\mu[\mathbf{g}](y).$$

Now, the model (6.1) can be written as

$$Y_x = \mathbf{g}^T(x)\boldsymbol{\theta} + \mathfrak{p}Z_x + \mathfrak{q}Z_x = \mathbf{g}^T(x)\boldsymbol{\theta}^{\mathfrak{q}} + \mathfrak{q}Z_x, \tag{6.12}$$

with $\boldsymbol{\theta}^{\mathfrak{q}} = \boldsymbol{\theta} + \mathbf{M}_{\mathbf{g}}^{-1}\int_{\mathscr{X}} \mathbf{g}(x)Z_x \mathrm{d}\mu(x)$. Since no informative prior on $\boldsymbol{\theta}$ is available, the prior on $\boldsymbol{\theta}^{\mathfrak{q}}$ is non-informative too (see Remark 6.2). The covariance kernel of $(\mathfrak{q}Z_x)_{x\in\mathscr{X}}$ in (6.12) is given by

$$K^{\mathfrak{q}}(x,y) = \mathsf{E}\big((\mathfrak{q}Z_x)(\mathfrak{q}Z_y)\big) = K(x,y) + \mathbf{g}^T(x)\mathbf{S}\mathbf{g}(y) - \mathbf{b}^T(x)\mathbf{g}(y) - \mathbf{g}^T(x)\mathbf{b}(y), \tag{6.13}$$

with $\mathbf{S} = \mathbf{M}_{\mathbf{g}}^{-1}\big(T_\mu[\mathbf{g}]\big|\mathbf{g}^T\big)_{L^2}\mathbf{M}_{\mathbf{g}}^{-1}$ and $\mathbf{b}(x) = \mathbf{M}_{\mathbf{g}}^{-1}T_\mu[\mathbf{g}](x)$. Such a kernel $K^{\mathfrak{q}}(\cdot,\cdot)$ is sometimes called a *reduction* of the kernel $K(\cdot,\cdot)$, see Schaback (1999).

Our motivation for introducing the model (6.12) is that we now have orthogonality in $L^2(\mathscr{X},\mu)$ between the realisations of $(\mathfrak{q}Z_x)_{x\in\mathscr{X}}$ and the trend subspace $\mathcal{T}$. The property below shows that predictions are not modified when using (6.12) instead of (6.1), i.e., when considering the kernel $K^{\mathfrak{q}}(\cdot,\cdot)$ instead of the kernel $K(\cdot,\cdot)$.

Theorem 6.1. *Assume that the design D_n is such that the design matrix $\mathbf{G}$ has full rank p (such a design is said to be $\mathcal{T}$-unisolvant). Then the two RF models* (6.1) *and* (6.12) *yield the same predictions and mean-squared prediction errors.*

This result is a direct consequence of the non-uniqueness of the kernels associated with a given semi-Hilbert space, and of the uniqueness of the optimal prediction in semi-Hilbert spaces, see Duchon (1977); Gauthier (2011). A proof is given in Appendix A. Notice that if we denote by $\widehat{Y}_x$ the resulting optimal linear prediction, and by $\widehat{\boldsymbol{\theta}^{\mathfrak{q}}}$ the underlying estimator of $\boldsymbol{\theta}^{\mathfrak{q}}$, then, by construction, the following orthogonality holds:

$$\int_{\mathscr{X}} \big(\widehat{Y}_x - \mathbf{g}^T(x)\widehat{\boldsymbol{\theta}^{\mathfrak{q}}}\big)\mathbf{g}(x)\mathrm{d}\mu(x) = 0.$$

Remark 6.1. The substitution of the kernel $K^q(\cdot,\cdot)$ for $K(\cdot,\cdot)$ can be related to the interpretation of the initial model (6.1) as an intrinsic random model, see Matheron (1973, 1971). For a real-valued function f on $\mathscr{X}$, let δ_x be the evaluation functional at $x \in \mathscr{X}$, that is $\delta_x[f] = f(x)$. Using a notation similar to Schaback (1999, Sect. 5), this kernel substitution amounts to replacing the evaluation functional δ_x by the functional $\delta_{(x)}$, defined by

$$\forall x \in \mathscr{X}, \quad f \in L^2(\mathscr{X},\mu), \quad \delta_{(x)}[f] = \mathfrak{q}f(x).$$

More precisely, for $g \in L^2(\mathscr{X},\mu)$, let $I_{g,\mu}$ denote the functional defined by $I_{g,\mu}[f] = (f|g)_{L^2}$. Using vector-matrix notation, we then have, for $x \in \mathscr{X}$, $\delta_{(x)} = \delta_x - \mathbf{g}^T(x)\mathbf{M}_{\mathbf{g}}^{-1}I_{\mathbf{g},\mu}$, and we can write $\delta_{(x)}[f] = \delta_x[\mathfrak{q}f] = ({}^t\mathfrak{q}\delta_x)[f]$. Notice that this kernel substitution is defined whenever $\mathcal{T} \subset L^2(\mathscr{X},\mu)$ and the assumptions of Section 2.2 are verified. In particular, it covers the case of a general linear trend $\mathbf{g}^T(x)\boldsymbol{\theta}$ in (6.1), whereas the theory of intrinsic random functions concerns translation-invariant kernels and only addresses the case of polynomial regression, see Matheron (1971). ◁

The integral operator associated with $K^{\mathfrak{q}}$ is

$$T_\mu^{\mathfrak{q}}[f](x) = \int_{\mathscr{X}} f(t)K^{\mathfrak{q}}(x,t)\mathrm{d}\mu(x),$$

with $f \in L^2(\mathscr{X},\mu)$ and $x \in \mathscr{X}$. By construction, it satisfies $T_\mu^{\mathfrak{q}}[g_j] = 0$ for all $j \in \{1,\cdots,p\}$. Denote by $\{\lambda_k^{\mathfrak{q}} | k \in \mathbb{I}_+^{\mathfrak{q}}\}$ the set of all strictly positive eigenvalues of $T_\mu^{\mathfrak{q}}$ and let $\varphi_k^{\mathfrak{q}}$ be their (canonically extended) associated eigenfunctions. We have

$$\tau^{\mathfrak{q}} = \int_{\mathscr{X}} K^{\mathfrak{q}}(x,x)\mathrm{d}\mu(x) = \tau - \mathrm{trace}\left(\mathbf{M}_{\mathbf{g}}^{-1}\left(T_\mu[\mathbf{g}]\middle|\mathbf{g}^T\right)_{L^2}\right) \leqslant \tau. \tag{6.14}$$

For a truncation subset $\mathbb{I}_{trc}^{\mathfrak{q}}$, using the same notations as in Section 6.2, the terms $D^{\mathfrak{q}}$ and $D_{trc}^{\mathfrak{q}}$ now equal 0, and inequalities similar to (2.5) are available, with $\tau_{err}^{\mathfrak{q}} = \sum_{k\notin\mathbb{I}_{trc}^{\mathfrak{q}}} \lambda_k^{\mathfrak{q}}$ quantifying the error due to spectral truncation.

Remark 6.2. Starting from the prior (6.2), we obtain

$$\mathsf{E}(\boldsymbol{\theta}^{\mathfrak{q}}) = \boldsymbol{\theta}_0 \text{ and } \mathsf{Cov}(\boldsymbol{\theta}^{\mathfrak{q}}) = \mathbf{A}^{\mathfrak{q}} = \mathbf{A} + \mathbf{S}.$$

When the prior is non-informative, that is, roughly speaking, when $\mathbf{A}^{-1} = 0$, then the same holds for $\mathbf{A}^{\mathfrak{q}}$ and the prior on $\boldsymbol{\theta}^{\mathfrak{q}}$ remains non-informative.

The situation would be different in presence of informative prior: in that case, orthogonality in $L^2(\Omega,\mathbb{P})$ between the components of $\boldsymbol{\theta}^{\mathfrak{q}}$ and the r.v. $\mathfrak{q}Z_x$, $x \in \mathscr{X}$, is lost when using the model (6.12), and $\mathsf{E}\big(\boldsymbol{\theta}^{\mathfrak{q}}(\mathfrak{q}Z_x)\big) = \mathbf{b}(x) - \mathbf{S}\mathbf{g}(x) = \mathbf{M}_{\mathbf{g}}^{-1}\mathfrak{q}\big(T_\mu[\mathbf{g}]\big)(x)$. The consequence on the IMSE calculation is that the matrix $\boldsymbol{\Gamma}^{\mathfrak{q}}$ corresponding to $\boldsymbol{\Gamma}$ in (6.6) is no longer block diagonal, with the two off-diagonal blocks being not trivial to evaluate. ◁

Remark 6.3. Comparing with (6.10), we now have $\left(\boldsymbol{\phi}_{trc}^{\mathfrak{q}}\middle|\mathbf{g}^T\right)_{L^2} = 0$. In order to further reduce the computational cost when using the reduced kernel $K^{\mathfrak{q}}(\cdot,\cdot)$, one may consider regressors $\mathbf{g}$ that form an orthonormal basis of $\mathcal{T}$ for $L^2(\mathscr{X},\mu)$, making $\mathbf{U}^{\mathfrak{q}}$ equal to the identity matrix. ◁

7 Numerical experiments

We consider a RF on $\mathscr{X} = [0,1]^d$, $d \in \mathbb{N}$, with kernel $K(x,y) = \prod_{i=1}^d K_{\ell_i}(x_i,y_i)$, where $x = (x_1,\cdots,x_d)$ and $K_{\ell_i}(x_i,y_i) = (1+\sqrt{3}|x_i-y_i|)\exp(\sqrt{3}|x_i-y_i|/\ell_i)$, $\ell_i > 0$ (Matérn 3/2). We use $d = 2$ in Sections 7.1, 7.2 and 7.4, and $d = 4$ in Section 7.3.

In all Section 7 we consider measures μ that correspond to pointwise quadrature approximations of the uniform probability on $[0,1]^d$, and we apply the methodology described in Section 5. Optimal design measures ν^* on the quadrature points are approximated by a vertex-exchange algorithm, see Böhning (1985, 1986). Quadrature points s_j that cannot be support points of the optimal measure can be removed from the search space using the criterion in (Pronzato, 2013). The iterations are stopped when the directional derivative (4.3) for the current approximated solution $\widehat{\nu}$ satisfies

$$\min_{j\in\{1,\ldots,N_q\}} F_{\Psi_\alpha}(\widehat{\nu},\delta_{s_j}) + \epsilon \geqslant 0, \tag{7.1}$$

which, by convexity, ensures that $\Psi_\alpha(\widehat{\nu}) \leqslant \Psi_\alpha(\nu^*) + \epsilon$, see (4.5). With a slight abuse of terminology, we refer to the obtained measures as the "optimal measures".

The IMSE-efficiency of a design D_n is measured by the ratio $\mathsf{IMSE}(D_n^*)/\,\mathsf{IMSE}(D_n)$, where D_n^* is the best n-point quadrature-design (i.e., a design only composed of quadrature-points) that we were able to obtain using the simulated-annealing algorithm presented in (Gauthier and Pronzato, 2015a). One should note that, although it is rather efficient, this global optimisation algorithm is much more time consuming than the convex-optimisation approach considered here. The model is without trend in Sections 7.1, 7.2 and 7.3; the presence of a linear parametric trend is considered in Section 7.4.

7.1 Regular grid approximation

Here μ is the discrete probability measure on $\mathscr{X} = [0,1]^2$ defined by a 33×33 regular square grid (midpoint rule), each grid-point receiving the same weight $1/N_q$, with $N_q = 33^2 = 1\,089$. We take $\ell_1 = \ell_2 = 0.15$.

Figure 1 shows the optimal measures for the heteroscedastic and homoscedastic models, respectively with $\sigma^2(x) = K_{err}(x,x)$ (left) and $\sigma^2 = \tau_{err}/\mu(\mathscr{X})$ (right), for $\alpha = n_{trc} = 7$ (so that $\tau_{trc} \approx 0.4484$). The optimal measure for the heteroscedastic model is supported by 11 points, but 97.71% of the mass is supported by 7 points only. The design $D_{7,het}^{supp}$ obtained by Algorithm 1 corresponds to those 7 points, with an IMSE-efficiency of 99.84%. A grid-restricted local descent starting from $D_{7,het}^{supp}$ converges to D_7^*. The optimal measure for the homoscedastic model is supported by 15 points, 7 of which (the same as above) carrying 79.39% of the total mass. The IMSE-efficiency of $D_{7,hom}^{supp}$ is 99.59%; $D_{7,hom}^{ext}$ obtained by local descent coincides also with D_7^*.

As illustrated by this first example, the optimal measures obtained with the heteroscedastic model have in general more support points than those obtained with the heteroscedastic model, which often complicates the extraction of a design with a given number of points n. Only the heteroscedastic model is used in the following.

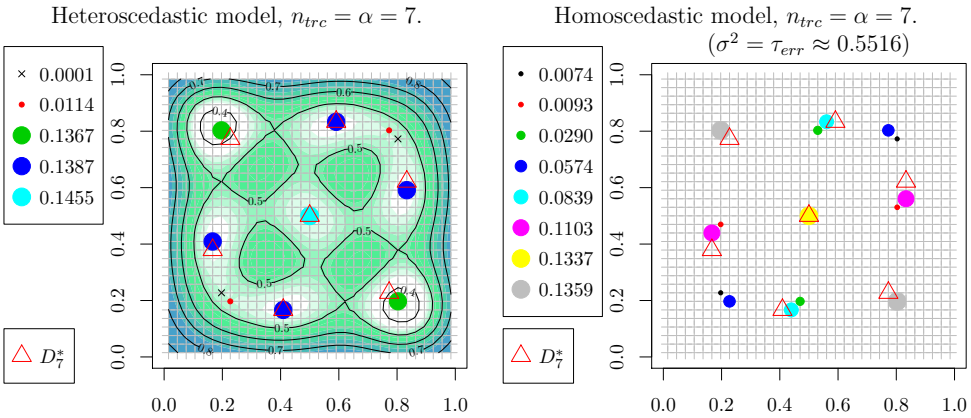

Figure 1: Contour-plot of the variance $x \mapsto \sigma^2(x) = K_{err}(x,x)$ ($\alpha = n_{trc} = 7$) and optimal measure ν_m^* for $\Psi_\alpha(\cdot)$ (disks with surface proportional to the weights p_k, $\epsilon =$ `1e-7` in (7.1)) for the heteroscedastic (left) and homoscedastic (right) models. The quadrature points are indicated by grey crosses, the IMSE-optimal 7-point quadrature-design is indicated by triangles.

Figure 2 shows the optimal measure ν_m^* for the heteroscedastic model, obtained for a fixed

truncation level $n_{trc} = 15$ and different values of α: the number of support points (or clusters of support points) of ν_m^* tends to increase when α increases. When $\alpha = 0.001$ (not shown), ν_m^* is reduced to a Dirac measure at the center of the grid. Remarkably, for this particular example the support of ν_m^* for $\alpha = 1$ coincides with the 13-point optimal quadrature-design D_{13}^*. The optimal measure for $\alpha = n_{trc} = 15$ is supported by 29 points, but 21 of them carry $96,75\%$ of the mass. Although the corresponding design D_{21}^{supp} is quite different from D_{21}^*, see Fig. 2-bottom-left, its IMSE efficiency is about 96.47% and a grid-restricted local descent yields a design D_{21}^{ext} with an IMSE-efficiency of 99.62%. The optimal measure for $\alpha = 100$ is supported by 37 points, and the design extraction procedure yields the same design D_{21}^{ext} as for $\alpha = 15$. More generally, we observe that for a given truncation level n_{trc}, the larger α, the more scattered the support of ν_m^*.

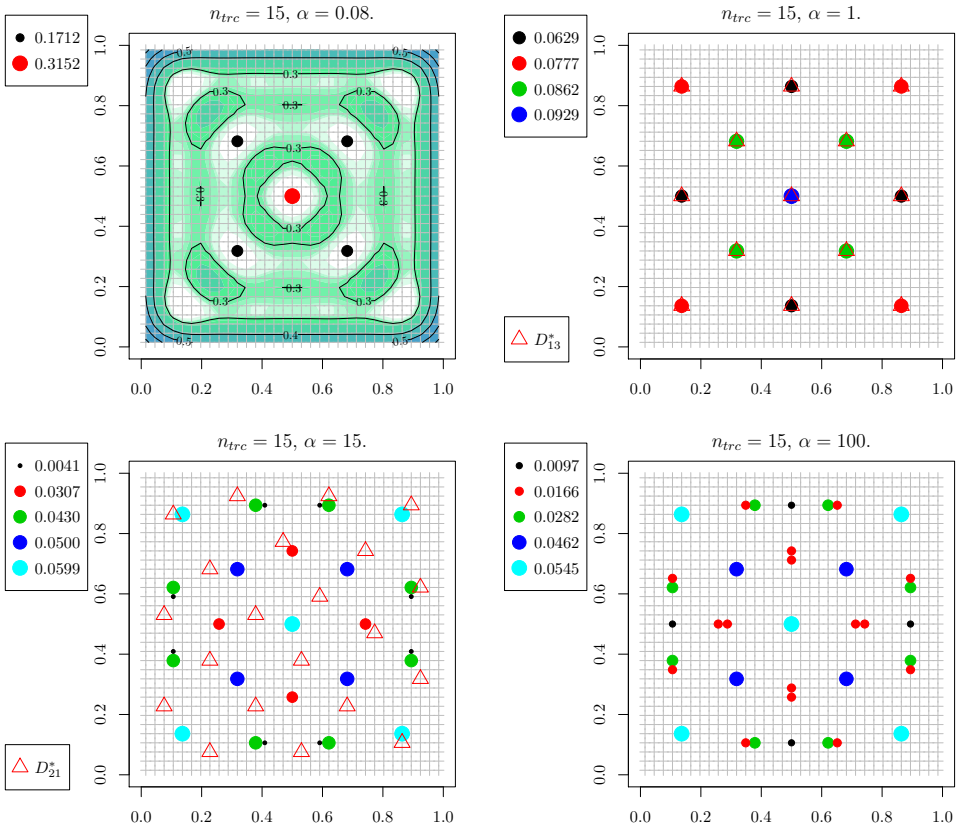

Figure 2: Optimal measures ($\epsilon =$ `1e-7` in (7.1)) for the heteroscedastic model with $n_{trc} = 15$ and $\alpha = 0.08$, 1, 15 and 100. A contour-plot of the variance $x \mapsto \sigma^2(x) = K_{err}(x, x)$ is also given (top-left).

Figure 3 presents the optimal measures obtained for $\alpha = 22$ and two different values of n_{trc}. For $n_{trc} = \alpha$ ($\tau_{trc} \approx 0.7648$), ν_m^* is supported by 52 points, among which 24 carry 78.96% of the mass. The optimal measure for $n_{trc} = 100$ ($\tau_{trc} \approx 0.9711$) is supported by 280 points, its exploitation for the extraction of a design with size $n \approx 22$ seems difficult. This motivates the recommandation of choosing $n_{trc} \approx \alpha \approx n$ to construct a design of size n.

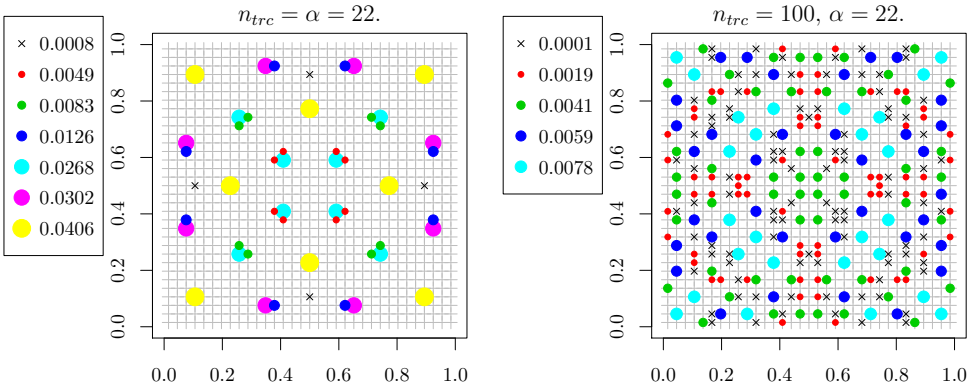

Figure 3: Optimal measures ($\epsilon =$ `1e-5` in (7.1)) for the heteroscedastic model with $\alpha = 22$ and $n_{trc} = 22$ (left), $n_{trc} = 100$ (right).

Figure 4 illustrates the behaviour of the design-extraction procedure (Algorithm 1) applied to the optimal measure for the heteroscedastic model with $n_{trc} = \alpha = 22$ (ν_m^* has $m = 52$ support points and is presented in Fig. 3-left). The evolution of ψ_k suggests the extraction of a design with size $n = 24$: we have $\psi_{m-24} - \psi_0 \approx$ `5.15e-4` and $\psi_{m-25} - \psi_0 \approx$ `7.11e-3`, and we can note a sudden increase of ψ_k when $k > m - 24$ (i.e., $n_s < 24$). The IMSE-efficiency of D_{24}^{supp} is about 99.37%, a quadrature-restricted local descent yields $D_{24}^{ext} = D_{24}^*$.

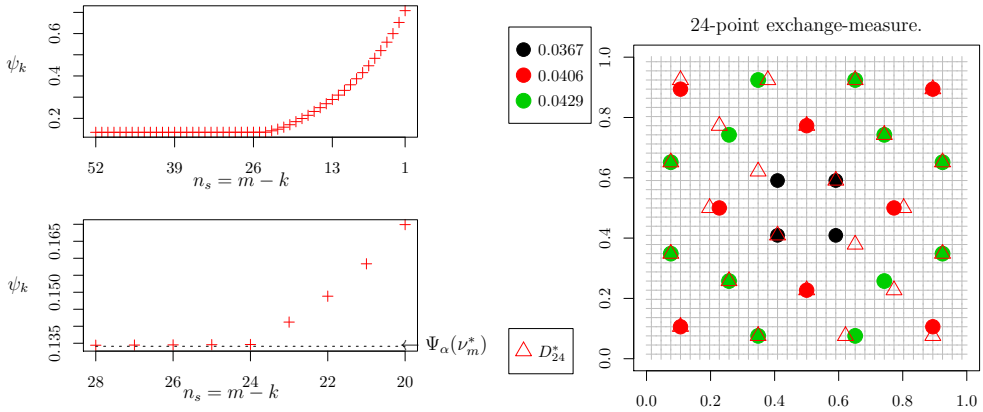

Figure 4: Design extraction for ν_m^* obtained with $n_{trc} = \alpha = 22$ (heteroscedastic model, ν_m^* is shown on Fig. 3-left): sequence $\{\psi_k\}$ (left) and measure ν_{24} obtained after $m - 24 = 28$ iterations of Algorithm 1 (right).

7.2 Quasi-Monte-Carlo quadrature

In order to illustrate the impact of the regularity of the quadrature on the optimal measures, μ corresponds now to a quadrature consisting of the $N_q = 1\,089$ first points of a low-discrepancy Halton sequence in $[0,1]^2$, all points receiving identical weights $1/N_q$.

Figure 5-left shows the optimal measure ν_m^* obtained for the heteroscedastic model with $n_{trc} = \alpha = 22$. There are 48 support points and ν_m^* presents some similarities with the measure presented in Fig. 3-left, which was obtained with a regular square grid. However, ν_m^* is now more irregular, as a consequence of dispersion of quadrature points (grey crosses) in the Halton sequence. Note in particular the presence of neighboring points with identical non-negligible weights in the rectangular box on the top of the figure, which may potentially confuse the design extraction.

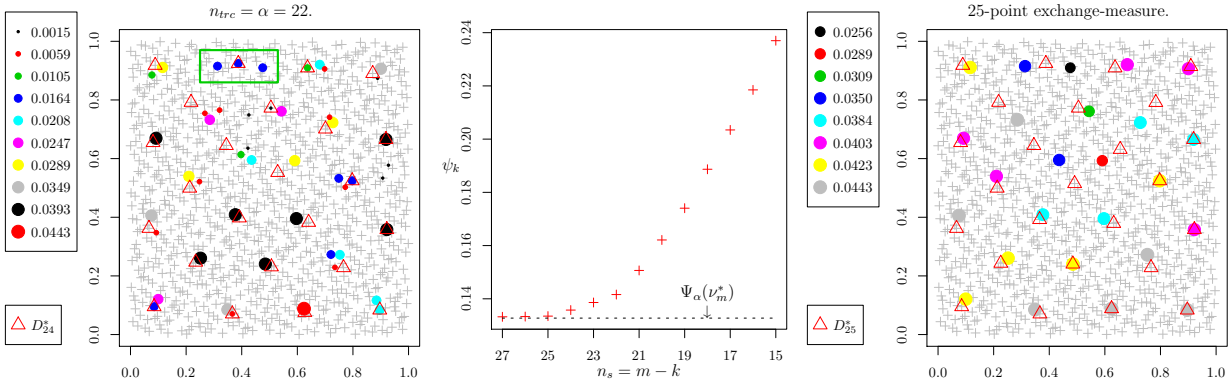

Figure 5: Optimal measure ($\epsilon =$ `1e-5` in (7.1)) for the heteroscedastic model with $\alpha = n_{trc} = 22$ on a low-discrepancy Halton sequence (left); sequence $\{\psi_k\}$ (middle); measure ν_{25} obtained after $m - 25 = 23$ iterations of Algorithm 1 (right).

When applying Algorithm 1, the sequence $\{\psi_k\}$ suggests the extraction of a design with size $n = 25$. Table 1 gives the IMSE-efficiencies obtained for D_n^{supp} and D_n^{ext} when the size n varies from 22 to 25. For $n = 23$ (respectively, $n = 24$), D_n^{ext} is an optimal (respectively, almost optimal) quadrature-design (in the sense that we were not able to obtain a better design on the quadrature points). For $n = 25$, the design-extraction procedure splits the cluster highlighted by a rectangle in Fig. 5-left into two design points, the efficiency of D_{25}^{ext} remaining reasonably high.

Table 1: IMSE-efficiencies (%) of D_n^{supp} and D_n^{ext} obtained from ν_m^* presented in Fig. 5-left.

n	22	23	24	25
D_n^{supp}	97.38	98.14	97.55	97.15
D_n^{ext}	99.94	100	100 ($-9\ 10^{-4}$)	99.44

7.3 Example with dimension $d = 4$

We take $d = 4$, with $\ell_1 = \ell_2 = \ell_3 = \ell_4 = 0.35$ in the $K_{\ell_i}(\cdot,\cdot)$. The measure μ corresponds to a $9 \times 9 \times 9 \times 9$ square grid (midpoint rule), all points receiving the same weight $1/N_q$, with $N_q = 9^4 = 6\,561$.

Figure 6 illustrates the results obtained for the heteroscastic model with $\alpha = n_{trc} = 31$ ($\tau_{trc} \approx 0.6658$). The optimal measure ν_m^* has 104 support points, among which 40 carry 94.74% of the mass. The sequence $\{\psi_k\}$ clearly suggests the extraction of a design of size $n = 40$. Remarkably enough, $D_{40}^{supp} = D_{40}^*$, but we must point out that such favourable situations are rather exceptional (on this example, not all truncation level n_{trc} lead to similar results).

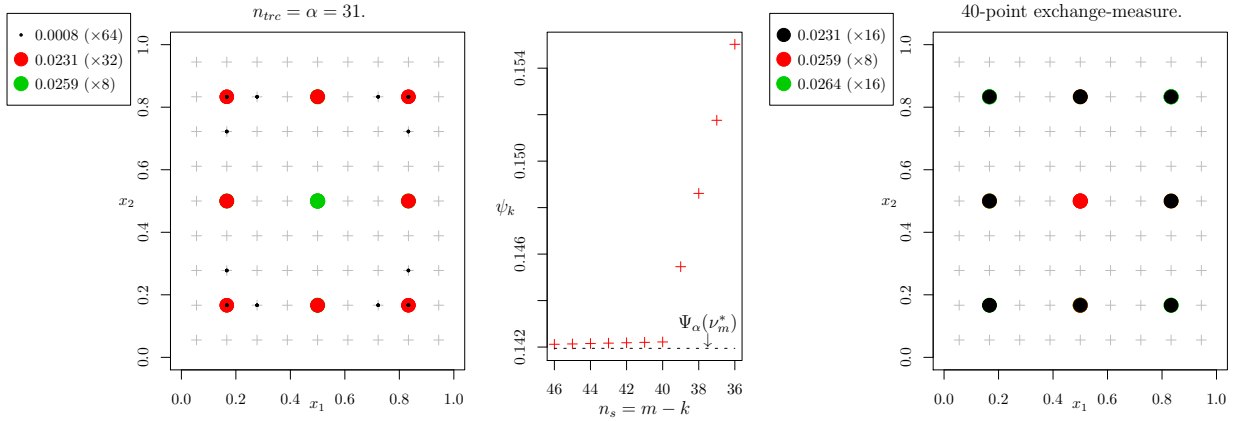

Figure 6: Projection on $\{x_1, x_2\}$ of the optimal measure ($\epsilon =$ `1e-6` in (7.1)) for the heteroscastic model with $\alpha = n_{trc} = 31$ (left), sequence $\{\psi_k\}$ (middle) and projection on $\{x_1, x_2\}$ of the measure ν_{40} obtained after $m - 40 = 64$ iterations of Algorithm 1 (right).

7.4 Models with unknown parametric trend

We consider the same framework as Section 7.1 but now assume the presence of a parametric trend, with $\mathbf{g}(x) = (g_1(x), g_2(x), g_3(x))^T = (1, x_1, x_2)^T$ for $x = (x_1, x_2) \in [0,1]^2$ ($p = 3$). We take $\mathsf{Cov}(\boldsymbol{\theta}) = \mathbf{A} = \mathrm{Id}_3$ when an informative prior on the trend-parameters is needed.

Figure 7-left illustrates the strong linear relationship that exists, in $L^2(\mathscr{X}, \mu)$, between the three trend-functions g_j and the eigenfunctions of the IMSE integral operator T_μ. This relationship can be interpreted as a kind of redundancy of the trend-functions in the models (6.1) or (6.5).

For a given truncation level n_{trc}, the total number of regressors n_{reg} in the BLM induced by the initial and reduced kernels is $n_{reg} = n_{trc} + p$, this corresponds to models of type (6.5). We have $n_{reg} = n_{trc}$ for the model, of type (3.1), induced by the augmented kernel. The middle and right parts of Fig. 7 aim at comparing the integrated variances of the error terms for the BLMs defined by the initial and modified kernels: initial kernel versus reduced kernel (middle), and initial kernel versus augmented kernel (right). We observe that for the same number of regressors n_{reg}, the reduced and augmented kernels yield BLMs that are more accurate than the BLM induced by

the initial kernel, in the sense that we have, for this particular example,

$$\tau^{\mathfrak{q}}_{err}[n_{trc}] \leqslant \tau_{err}[n_{trc}] \text{ and } \tau^{\text{full}}_{err}[n_{trc}+p] \leqslant \tau_{err}[n_{trc}],$$

where $[n_{trc}]$ and $[n_{trc}+p]$ indicate the number of eigenfunctions considered (we recall that $\tau_{err} = \int_{\mathscr{X}} K_{err}(x,x)\mathrm{d}\mu(x)$).

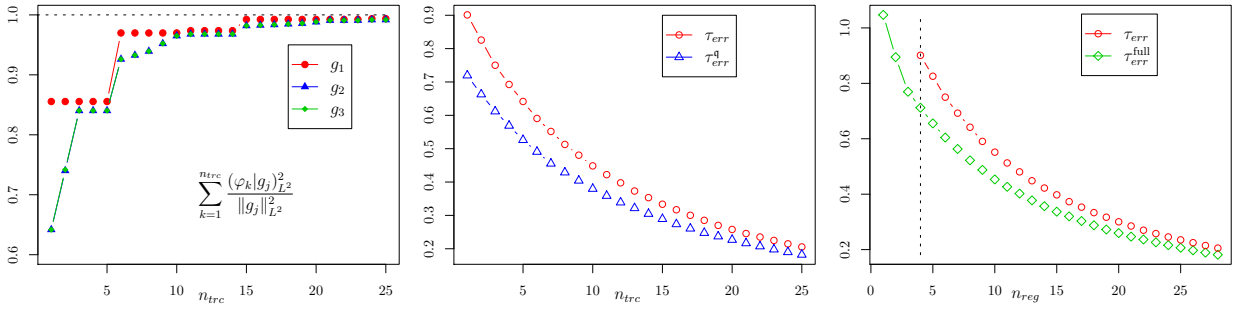

Figure 7: Contributions of the first n_{trc} eigenfunctions φ_k to the trend-functions g_1, g_2 and g_3 (left), and values of τ_{err}, $\tau^{\mathfrak{q}}_{err}$ and τ^{full}_{err} as functions of the truncation level (middle and right).

Figure 8 gives a further illustration that, for the same number of regressors considered, the BLM induced by the reduced kernel contains more information than the BLM induced by the initial kernel. The optimal measures ν^*_m for the heteroscedastic models are presented, on the left for the initial kernel with $n_{trc} = \alpha = 22$ (and therefore 25 regressors), on the right for the reduced kernel with $n_{trc} = 19$ (22 regressors), both with $\alpha = 22$ (no prior on $\boldsymbol{\theta}$ is used). The two measures look rather similar, and the contour plots of the variances $x \mapsto K_{err}(x,x)$ and $x \mapsto K^{\mathfrak{q}}_{err}(x,x)$ also present strong similitude (with $\tau_{err} = 0.2352$ and $\tau^{\mathfrak{q}}_{err} = 0.2370$). It thus appears that the BLM induced by the initial kernel requires 25 regressors to carry similar information as the BLM induced by the reduced kernel with 22 regressors only. Also, ν^*_m is supported by 56 points for the initial kernel, and has only 48 support points for the reduced kernel. Algorithm 1 applied to both measures yields the same 24-point design D^{ext}_{24}, which coincides with D^*_{24}.

8 Concluding discussion

We have shown that the convex relaxation method described by Fedorov (1996); Spöck and Pilz (2010) can be efficiently applied when considering the IMSE-adapted Karhunen-Loève expansion of the RF. The obtained Bayesian A-optimality criterion $\Psi_\alpha(\cdot)$ is closely related to the truncated-IMSE criterion for RF interpolation models considered for instance in (Gauthier and Pronzato, 2014). A numerical implementation of the approach has been proposed, based on a quadrature approximation of the IMSE with restriction to designs supported by quadrature points. Efficient convex-programming algorithms can then be used to construct IMSE-optimal design measures, with guaranteed convergence to the optimum.

A greedy exchange algorithm has been presented for the extraction of an exact design D_n of given size n from an optimal measure ν^*_m, and the adaptation of ν^*_m to n via the tuning of the parameters α and n_{trc} that enter $\Psi_\alpha(\cdot)$ has been discussed. We have observed that the heteroscedastic approximate

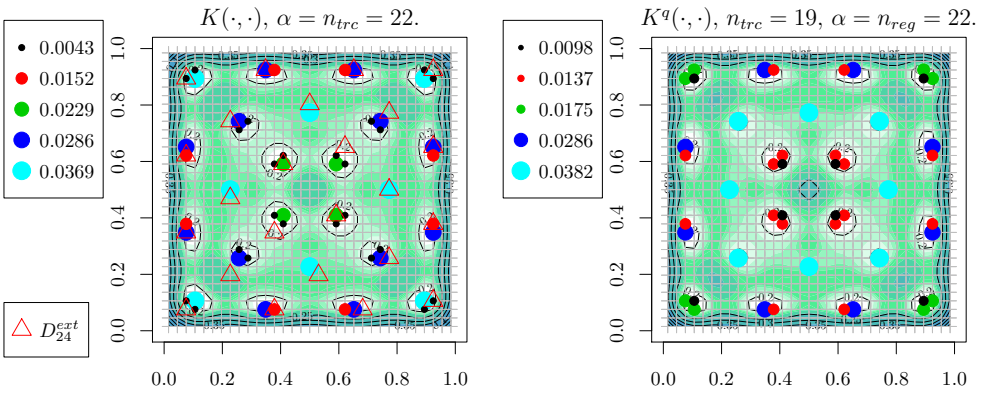

Figure 8: Optimal measures for the heteroscedastic models ($\epsilon =$ `1e-5` in (7.1)) induced by the initial kernel $K(\cdot,\cdot)$ with $n_{trc} = \alpha = 22$ (left), and by the reduced kernel $K^{\mathfrak{q}}(\cdot,\cdot)$ with $n_{trc} = 19$ and $\alpha = 22$ (right).

BLM, with $\sigma^2(x) = K_{err}(x,x)$, often yields optimal measures that are more concentrated than the ones obtained with the homoscedastic model, and are thus easier to exploit for the extraction of exact designs of given size. In line with the numerical results presented in (Gauthier and Pronzato, 2015a), we also observed that when α and n_{trc} are such that the size n of the design D_n extracted is close to n_{trc}, D_n inherits a high IMSE-efficiency. Our numerical experiments indicate that it is possible to modulate the size of the designs extracted by considering different truncation levels n_{trc}, and that values $n_{trc} \approx \alpha \approx n$ are recommended. Further developments might help to select more precisely values of α and n_{trc} especially suited for the extraction of a design of size n. The construction of other design-extraction procedures also deserves further investigations.

We have proposed two extensions to the direct approach of Spöck and Pilz (2010) for RF models that include a linear parametric trend. One is based on kernel augmentation, and applies to the case where an informative prior on the two first moments of the trend-parameters $\boldsymbol{\theta}$ is available. The other amounts to a kernel reduction defined from a linear continuous projection onto a subspace of the trend-space, and is for the case when no such prior information on $\boldsymbol{\theta}$ is available. They permit in particular to bound the error induced by considering the truncated-IMSE instead of the true IMSE, whereas the derivation of such error bounds seems much more difficult with the initial kernel. The numerical experiments carried out in Section 7.4 also point out that, for an equivalent number of regressors, the modified kernels generally lead to BLMs having smaller errors than the BLMs induced by the initial kernel. Theorem 6.1 shows the equivalence between the predictions induced by the initial model and those with a model based on the reduced kernel in absence of informative prior on $\boldsymbol{\theta}$. This is a very general result, with potential consequences in other contexts involving RF models.

Finally, only Bayesian A-optimality has been considered, due to its direct connection with the truncated-IMSE for the initial RF model. Other choices are possible and deserve further investigations, see in particular Fedorov (1996); Spöck and Pilz (2010) for Bayesian D-optimality.

Acknowledgements

This work was partially supported by the ANR project DESIRE (DESIgns for spatial Random fiElds), nb. 2011-IS01-001-01, joint with the Statistics Department of the Johannes Kepler Universität, Linz (Austria). B. Gauthier also wishes to thanks the MRI department of EdF lab, Chatou, France.

A Proofs

Proof of Proposition 6.1. Consider the Cholesky decomposition $\mathbf{K} = \mathbf{C}\mathbf{C}^T$. For $k \in \mathbb{I}_+$, let $\boldsymbol{\phi}_k = \big(\varphi_k(x_1), \cdots, \varphi_k(x_n)\big)^T$. Using developments similar to those used to obtain (3.6), we have

$$C - C_{trc} = \sum_{k \notin \mathbb{I}_{trc}} \lambda_k (\sqrt{\lambda_k}\mathbf{C}^{-1}\boldsymbol{\phi}_k)^T \big(\mathrm{Id} - \mathbf{C}^{-1}\mathbf{G}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{C}^{-T}\big)(\sqrt{\lambda_k}\mathbf{C}^{-1}\boldsymbol{\phi}_k).$$

From Gauthier and Pronzato (2014), we know that, for all $k \in \mathbb{I}_{trc}$,

$$0 \leqslant (\sqrt{\lambda_k}\mathbf{C}^{-1}\boldsymbol{\phi}_k)^T(\sqrt{\lambda_k}\mathbf{C}^{-1}\boldsymbol{\phi}_k) \leqslant 1. \tag{A.1}$$

Consider the matrix $\mathbf{Q} = \mathbf{C}^{-1}\mathbf{G}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{C}^{-T}$; it satisfies $\mathbf{Q} = \mathbf{Q}^T$ and

$$\mathbf{Q}^2 = \mathbf{Q} - \mathbf{C}^{-1}\mathbf{G}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{A}^{-1}\big(\mathbf{G}^T\mathbf{K}^{-1}\mathbf{G} + \mathbf{A}^{-1}\big)^{-1}\mathbf{G}^T\mathbf{C}^{-T}. \tag{A.2}$$

The second matrix on the right-hand side of (A.2) is symmetric and non-negative definite. Therefore, $\mathbf{Q}^2 \preceq \mathbf{Q}$ (Loewner ordering), and similarly $(\mathrm{Id}-\mathbf{Q})^2 \preceq (\mathrm{Id}-\mathbf{Q})$. The real matrix $(\mathrm{Id}-\mathbf{Q})$ is therefore positive and contractant, which, combined with (A.1), concludes the proof. □

Proof of Theorem 6.1. For $x \in \mathscr{X}$, let $\mathbf{c}_x^T\mathbf{y}$ be a linear prediction of Y_x for the model (6.1) with no informative prior on $\boldsymbol{\theta}$. The MSE associated with this prediction is given by

$$s^2(x) = \mathsf{E}\big((Y_x - \mathbf{c}_x^T\mathbf{y}^2)\big) = \big((\mathbf{g}(x) - \mathbf{G}^T\mathbf{c}_x)^T\boldsymbol{\theta}\big)^2 + K(x,x) + \mathbf{c}_x^T\mathbf{K}\mathbf{c}_x - 2\mathbf{c}_x^T\mathbf{k}(x).$$

The no-bias condition implies $\mathbf{G}^T\mathbf{c}_x = \mathbf{g}(x)$. The stationary condition for the Lagrangian (a necessary and sufficient condition for the minimisation of $s^2(x)$ with respect to $\mathbf{c}_x$ under the no-bias constraint) is

$$\begin{pmatrix} 0 & \mathbf{G}^T \\ \mathbf{G} & \mathbf{K} \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda} \\ \mathbf{c}_x \end{pmatrix} = \begin{pmatrix} \mathbf{g}(x) \\ \mathbf{k}(x) \end{pmatrix}. \tag{A.3}$$

If we consider the model (6.12) and a prediction of the form $(\mathbf{c}_x^{\mathfrak{q}})^T\mathbf{y}$, we obtain an equation of the same type as (A.3), where $\mathbf{c}_x$, $\boldsymbol{\lambda}$, $\mathbf{K}$ and $\mathbf{k}(x)$ are replaced by $\mathbf{c}_x^{\mathfrak{q}}$, $\boldsymbol{\lambda}^{\mathfrak{q}}$, $\mathbf{K}^{\mathfrak{q}}$ and $\mathbf{k}^{\mathfrak{q}}(x)$ respectively. Applying the no-bias condition, we finally obtain

$$\begin{pmatrix} 0 & \mathbf{G}^T \\ \mathbf{G} & \mathbf{K} - \mathbf{G}\mathbf{B}^T \end{pmatrix} \begin{pmatrix} \boldsymbol{\lambda}^{\mathfrak{q}} \\ \mathbf{c}_x^{\mathfrak{q}} \end{pmatrix} = \begin{pmatrix} \mathbf{g}(x) \\ \mathbf{k}(x) - \mathbf{G}\mathbf{b}(x) \end{pmatrix}, \tag{A.4}$$

with $\mathbf{B}^T = \big(\mathbf{b}(x_1), \cdots, \mathbf{b}(x_n)\big)$. Therefore, if $(\boldsymbol{\lambda}, \mathbf{c}_x)$ satisfies (A.3), then $\mathbf{c}_x^{\mathfrak{q}} = \mathbf{c}_x$ and $\boldsymbol{\lambda}^{\mathfrak{q}} = \boldsymbol{\lambda} + \mathbf{B}^T\mathbf{c}_x - \mathbf{b}(x)$ are solution of (A.4). The two optimal linear predictions $(\mathbf{c}_x)^T\mathbf{y}$ and $(\mathbf{c}_x^{\mathfrak{q}})^T\mathbf{y}$ thus coincide and one can check that $s_{\mathfrak{q}}^2(x) - s^2(x) = 0$. □

References

Barvinok, A. (2002). *A Course in Convexity*, volume 54 of *Graduate Studies in Mathematics*. American Mathematical Society, Providence, Rhode Island.

Böhning, D. (1985). Numerical estimation of a probability measure. *Journal of Statistical Planning and Inference*, 11(1):57–69.

Böhning, D. (1986). A vertex-exchange-method in D-optimal design theory. *Metrika*, 33(1):337–347.

Chaloner, K. (1984). Optimal Bayesian experimental design for linear models. *The Annals of Statistics*, pages 283–300.

Duchon, J. (1977). Splines minimizing rotation-invariant semi-norms in Sobolev spaces. In Schempp, W. and Zeller, K., editors, *Constructive Theory of Functions of Several Variables*, pages 85–100. Springer, Berlin.

Fang, K.-T., Li, R., and Sudjianto, A. (2010). *Design and Modeling for Computer Experiments*. CRC Press, Boca Raton.

Fedorov, V. (1972). *Theory of Optimal Experiments*. Academic Press, New York.

Fedorov, V. (1996). Design of spatial experiments: model fitting and prediction. In Gosh, S. and Rao, C., editors, *Handbook of Statistics, vol. 13*, chapter 16, pages 515–553. Elsevier, Amsterdam.

Gauthier, B. (2011). *Approche spectrale pour l'interpolation à noyaux et positivité conditionnelle*. PhD thesis, Ecole Nationale Supérieure des Mines de Saint-Etienne.

Gauthier, B. and Pronzato, L. (2014). Spectral approximation of the IMSE criterion for optimal designs in kernel-based interpolation models. *SIAM/ASA Journal on Uncertainty Quantification*, 2:805–825.

Gauthier, B. and Pronzato, L. (2015a). Approximation of IMSE-optimal designs via quadrature rules and spectral decomposition. *Communications in Statistics – Simulation and Computation*. To appear, DOI:10.1080/03610918.2014.972518, `http://hal.archives-ouvertes.fr/hal-00936681`.

Gauthier, B. and Pronzato, L. (2015b). Optimal design for prediction in random field models via covariance kernel expansions. In J., C. M. and Atkinson, A., editors, *mODa'11 – Advances in Model-Oriented Design and Analysis, Proceedings of the 11th Int. Workshop, Hamminkeln-Dingden (Germany)*, Heidelberg. Physica Verlag. to appear.

Harari, O. and Steinberg, D. (2014). Optimal designs for Gaussian process models via spectral decomposition. *Journal of Statistical Planning and Inference*, 154:87–101.

Matheron, G. (1971). La théorie des fonctions aléatoires intrinsèques généralisées. *Note no. 117 du Centre de Géostatistique de l'Ecole des Mines de Paris*.

Matheron, G. (1973). The intrinsic random functions and their applications. *Advances in Applied Probability*, pages 439–468.

Omre, H. and Halvorsen, K. (1989). The Bayesian bridge between simple and universal kriging. *Mathematical Geology*, 21(7):767–786.

Pilz, J. (1983). *Bayesian Estimation and Experimental Design in Linear Regression Models*, volume 55. Teubner-Texte zur Mathematik, Leipzig. (also Wiley, New York, 1991).

Pronzato, L. (2013). A delimitation of the support of optimal designs for Kiefer's ϕ_p-class of criteria. *Statistics & Probability Letters*, 83(12):2721–2728.

Pronzato, L. and Pázman, A. (2013). *Design of Experiments in Nonlinear Models. Asymptotic Normality, Optimality Criteria and Small-Sample Properties.* Springer, LNS 212, New York, Heidelberg.

Pukelsheim, F. (1993). *Optimal Experimental Design.* Wiley, New York.

Pukelsheim, F. and Reider, S. (1992). Efficient rounding of approximate designs. *Biometrika*, 79(4):763–770.

Rasmussen, C. and Williams, C. (2006). *Gaussian Processes for Machine Learning.* MIT press, Cambridge, MA.

Sacks, J., Welch, W., Mitchell, T., and Wynn, H. (1989). Design and analysis of computer experiments. *Statistical Science*, 4(4):409–423.

Santner, T., Williams, B., and Notz, W. (2003). *The Design and Analysis of Computer Experiments.* Springer, New York.

Schaback, R. (1999). Native Hilbert spaces for radial basis functions I. In *New Developments in Approximation Theory*, pages 255–282. Springer.

Spöck, G. and Pilz, J. (2010). Spatial sampling design and covariance-robust minimax prediction based on convex design ideas. *Stochastic Environmental Research and Risk Assessment*, 24(3):463–482.

Winkler, G. (1988). Extreme points of moment sets. *Mathematics of Operations Research*, 13(4):581–587.